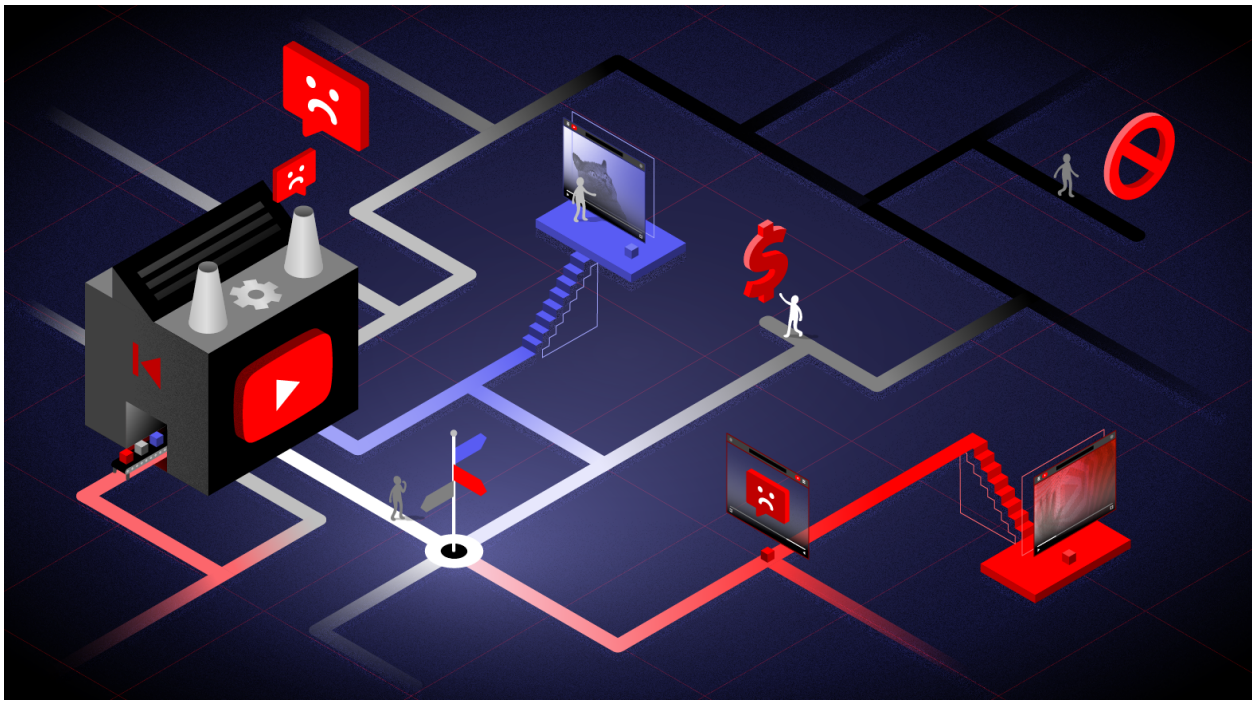
moz://a

YouTube Regrets

A crowdsourced investigation into YouTube's recommendation algorithm



Juli 2021

Inhaltsangabe

Kurzfassung

YouTube ist die Plattform mit den zweitmeisten Besuchen weltweit und der YouTube-Algorithmus trägt zu 70 % der Zeit bei, die Nutzer:innen mit dem Ansehen von Videos verbringen – ungefähr 700 Millionen Stunden[1] jeden Tag. Jahrelang hat der Empfehlungs-Algorithmus bei der Verbreitung von Fehlinformationen und Desinformationen zu Themen wie Gesundheit und Politik, Hasstiraden und anderen bedauerlichen Inhalten auf der ganzen Welt beigetragen. Aufgrund des enormen Einflusses von YouTube erreichen diese Filme ein riesiges Publikum und nehmen schwerwiegenden Einfluss auf unzählige Leben, von Radikalisierung bis Polarisierung.[2] Trotzdem reagiert YouTube auf Kritik mit Untätigkeit und Undurchsichtigkeit.

Nach dem jahrelangen Einsatz Mozillas für mehr Transparenz bei YouTubes Empfehlungs-Algorithmus und der Forderung, Forscher:innen die Plattform untersuchen zu lassen, hat Mozilla 2020 auf die Untätigkeit der Plattform reagiert, indem es die Nutzer:innen dazu befähigte, die Misstände anzusprechen. Wir brachten RegretsReporter auf den Weg, eine Browser-Erweiterung und ein durch Crowdsourcing angetriebenes Projekt, das dabei helfen soll, die Schäden, die YouTubes Algorithmus bei Nutzer:innen anrichten kann, besser zu verstehen.

37.380 YouTube-Nutzer:innen meldeten sich zum Überwachungsdienst und stellten uns freiwillig Daten über bedauerliche Erfahrungen zur Verfügung, die sie bei YouTube gemacht haben, damit unsere Forscher:innen sie genau analysieren konnten. Das Resultat: Mozilla erhielt in der bis dato größten, jemals durch Crowdsourcing angetriebenen Untersuchung von YouTubes Algorithmus Einblick in eine Datensammlung von Daten, die YouTube sonst gut abschirmt. Insgesamt meldeten die freiwilligen Helfer:innen zwischen Juli 2020 und Mai 2021 3.362 Videos aus 91 Ländern.

---

[1] YouTube gibt an, dass täglich Videos gespielt werden, die zusammengenommen rund eine Milliarde Stunden lang sind. Bei der CES 2018 erläuterte Neal Mohan, YouTubes CPO, dass mindestens 70 % dieser Zeit auf Empfehlungen zurückzuführen seien, die von der KI generiert werden. Soweit diese Zahlen momentan korrekt sind, sind KI-generierte Empfehlungen auf YouTube für 700 Millionen Stunden an Videoansichten verantwortlich.

[2] Untersuchungen von Mozilla, der Anti-Defamation League, der New York Times, der Washington Post, des Wall Street Journal und einiger weiterer Organisationen, akademischer Forscher:innen und Verlage haben offengelegt, wie YouTube Nutzer:innen fehlinformieren, polarisieren und radikalisieren kann.

Dieser Bericht fasst zusammen, was wir aus der RegretsReporter-Recherche gelernt haben. Wir haben hauptsächlich drei Dinge feststellen können:

- **YouTube Regrets sehen ganz unterschiedlich aus und sind verstörend.** Unsere freiwilligen Helfer:innen meldeten alles von Panikmache rund ums Thema COVID-19 über die Verbreitung politischer Fehlinformationen bis hin zu völlig unangemessenen „Kinder“-Cartoons. Die am häufigsten auftretenden Kategorien an YouTube Regrets sind Fehlinformationen, brutale oder verstörende Inhalte, Hasstiraden und Spam/Betrug.

- **Der Fehler liegt im Algorithmus.** 71 % aller gemeldeten Regrets stammten aus Videos, die YouTubes automatisches Empfehlungssystem unseren freiwilligen Helfer:innen vorgeschlagen hatte. Zudem wurden empfohlene Videos mit 40 % höherer Wahrscheinlichkeit gemeldet als Videos, nach denen aktiv gesucht wurde. Und in mehreren Fällen empfahl YouTube sogar Videos, die gegen die eigenen Community-Richtlinien der Plattform verstoßen und/oder gar keinen Bezug zu den vorher angesehenen Videos hatten.

- **Am stärksten trifft es Nicht-Englischsprechende.** In Ländern, deren Primärsprache nicht Englisch ist, ist der Anteil an YouTube Regrets ganze 60 % höher als in englischsprachigen Ländern. In Brasilien, Deutschland und Frankreich ist sie besonders hoch. Auch pandemiebezogene Regrets kamen besonders häufig in nicht-englischsprachigen Ländern vor.

In diesem Bericht schauen wir uns die Ergebnisse anhand Datenanalysen und umfangreichen Fallstudien genau an. Zur Veranschaulichung unserer Untersuchung reichen bereits einige wenige Beispiele. Ein:e freiwillige:r Helfer:in meldete das animierte Video „Woody's Got Wood“, eine sexualisierte Parodie des Kinderfilms „Toy Story“. Jemand anderem wurde eine widerlegte Verschwörungstheorie über die amerikanische Aktivistin und Politikerin Ilhan Omar und Wahlbetrug vorgeschlagen. Wieder jemand anderem wurde ein Video mit dem Namen „Blacks in Power Don't Empower Blacks“ (auf Deutsch: Schwarze in Machtposition ermächtigen nicht andere Schwarze), in dem rassistische und beleidigende Sprache und Ideen geäußert werden.

Mozilla will diese Probleme nicht nur diagnostizieren – sondern zu einer Lösung beitragen. Im Abschnitt „Empfehlungen“ dieses Berichts befindet sich eine klare

Anleitung für YouTube, andere Internet-Plattformen, Gesetzgeber und die Öffentlichkeit. Zu diesen Empfehlungen zählen unter anderem:

- **Plattformen müssen** es Forscher:innen ermöglichen, die Systeme für Empfehlungen zu prüfen.
- **Plattformen müssen** Informationen darüber offenlegen, wie ihre Systeme für Empfehlungen funktionieren, und transparente Berichte erstellen, die einen ausreichenden Einblick in Problemgebiete und den Fortschritt auf diesen Gebieten gewähren.
- **Gesetzgeber müssen** von YouTube verlangen, Informationen freizugeben und Tools zu entwerfen, die eine unabhängige Auseinandersetzung mit YouTubes Empfehlungsalgorithmen ermöglichen.
- **Gesetzgeber müssen** Forscher:innen, Journalist:innen und andere Überwacher:innen schützen, die alternative Methoden nutzen als die durch die Plattformen bereitgestellten, um sie zu untersuchen.
- **Menschen sollten** ihre Dateneinstellungen auf YouTube und Google aktualisieren und sicherstellen, dass sie sich und ihre Familie ausreichend schützen.

Außerdem enthält dieser Bericht einen umfangreichen Abschnitt über die Methodologie, der unsere Untersuchung beschreibt. Darin wird im Detail auf die Fragen eingegangen, die wir für unsere Untersuchung gestellt haben, die Funktionsweise unserer Browser-Erweiterung und unsere Analysemethoden beschrieben und mehr.

Einleitung

Sam war ein 13-jähriger Junge, der an Depressionen litt und der in dem Moment, in dem er am verletzlichsten war rekrutiert wurde, der Alt-Right-Bewegung beizutreten. Sams Radikalisierung ist zum Teil darauf zurückzuführen, dass er immer tiefer in das Labyrinth empfohlener Inhalte auf YouTube und anderen Sozialen Medien geriet. In einem Interview mit Mozilla erklärte Sams Mutter, dass „Plattformen wie YouTube und andere scheinbar alles daransetzen, keine Rechenschaft abzulegen. Die Unternehmen verstecken sich hinter Argumenten wie Meinungsfreiheit, aber mir kam das immer vor wie ein praktischer Weg, keine Verantwortung zu übernehmen und mehr Geld reinzubekommen. Kinder sind besonders verletzlich und ganz besonders Kinder, die im Internet nach Informationen und Unterstützung suchen."

Sams Story ist kein Einzelfall. Forscher:innen und investigative Journalist:innen dokumentieren seit Jahren, wie YouTube-Empfehlungen Menschen in die dunkelsten und extremsten Ecken des Internets führen können. YouTube ist natürlich nicht *allein* für komplexe Probleme wie Radikalisierung verantwortlich, allerdings gibt es Beweise, die nahelegen, dass der Empfehlungsalgorithmus des Unternehmens eine übermächtige Rolle dabei spielt. Der Algorithmus steuert Nutzer:innen auf Inhalte zu und zeigt ihnen, sobald sie sich dann erst mal im Labyrinth befinden, mehr und mehr immer radikalere Ideen.

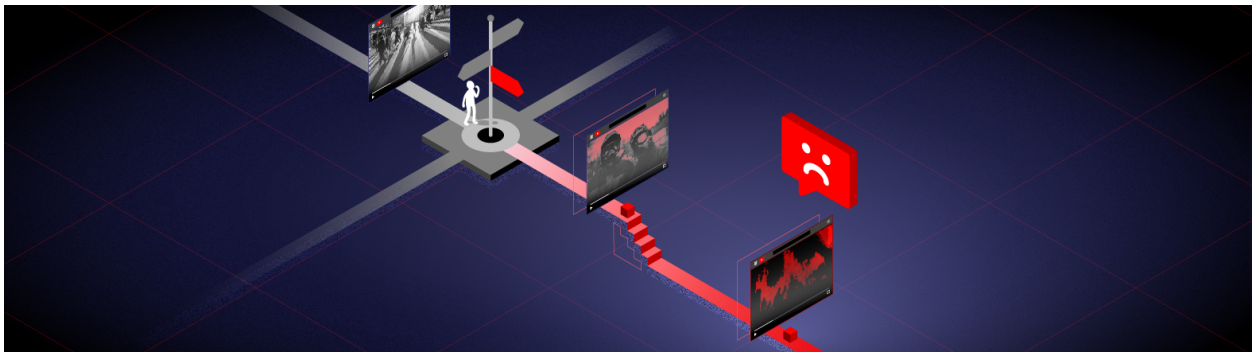
Dieses Problem mit YouTubes Algorithmus ist Teil eines größeren Problems: des undurchsichtigen mysteriösen Einflusses, den kommerzielle Algorithmen auf unsere Leben nehmen können. **YouTubes Algorithmus generiert geschätzte 700 Millionen Stunden an Videoansichten jeden Tag und trotzdem weiß die Öffentlichkeit nur sehr wenig darüber, wie er funktioniert.** Wir haben keine Möglichkeit, den Algorithmus zu untersuchen, und wenn Storys wie die von Sam an die Öffentlichkeit geraten, ist es unmöglich herauszufinden, was genau passiert ist und wie wir verhindern können, dass es in Zukunft wieder passiert. Im Kern arbeitet YouTubes Algorithmus im Interesse von YouTube und nicht dem der Öffentlichkeit.

Seit 2019 setzt sich Mozilla für Transparenz im Bezug auf YouTubes Algorithmus ein. Unsere Kampagne YouTube Regrets hebt 28 Geschichten wie die von Sam hervor – von Menschen, deren Leben durch ihre Erlebnisse auf YouTube auf die schiefe Bahn gerieten und in einigen Fällen für immer verändert wurden. Wir haben diese

Geschichten veröffentlicht, um auf den Einfluss von nicht vertrauenswürdigen Algorithmen auf Menschenleben aufmerksam zu machen und um YouTube dazu zu bewegen, mehr Verantwortung dafür zu übernehmen und offener darüber zu sein, wie der Algorithmus der Plattform funktioniert. Zwei Jahre später warten wir immer noch darauf, dass YouTube etwas unternimmt.

Deshalb haben wir den RegretsReporter entwickelt. RegretsReporter ist ein Crowdsourcing-Tool, mit dessen Hilfe wir den Umfang der ursprünglichen YouTube Regrets-Kampagne erweitern können. Das Tool ermöglicht es Mozillas freiwilligen Helfer:innen, Daten über ihre Erfahrungen auf YouTube beizutragen. RegretsReporter ähnelt anderen Browsererweiterungen, die mit alternativen Methoden einige der Algorithmen untersuchen, die den größten Einfluss auf die Öffentlichkeit nehmen. Dabei sind diese Tools kein Ersatz für echte Transparenz für Menschen und Institutionen, aber sie sind unsere beste Option, um Unternehmen für ihren Einfluss auf das Leben der Menschen zur Verantwortung zu ziehen.

Mit dieser Untersuchung wollen wir die Ergebnisse teilen, die wir mithilfe von RegretsReporter erarbeitet haben. Dabei hoffen wir, dass diese Ergebnisse – die wahrscheinlich nur die Spitze des Eisbergs sind – die Öffentlichkeit und diejenigen, die im Dienst der Öffentlichkeit handeln, davon überzeugen, dass wir unbedingt volle Transparenz für YouTubes Algorithmus brauchen.



Was meinen wir mit 'Regret'?

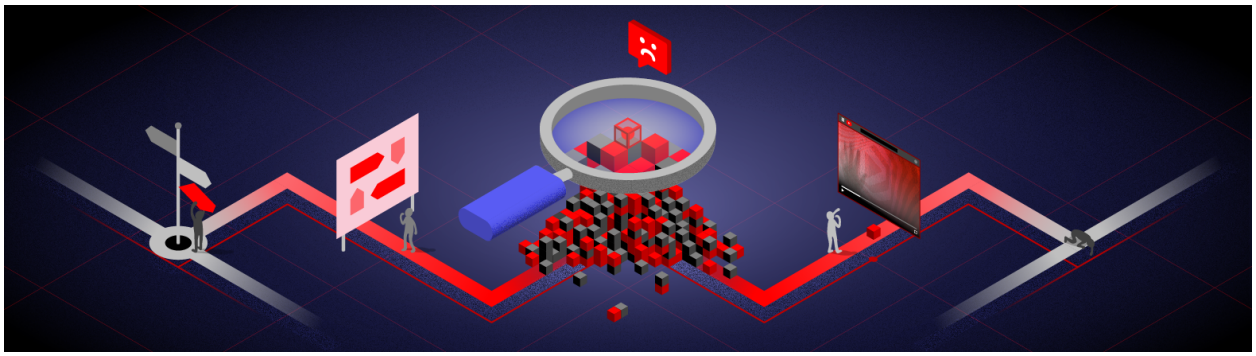
Das Konzept „YouTube Regret“ wurde aus einer globalen, mittels Crowdsourcing angetriebenen Kampagne geboren, die Mozilla 2019 entwickelte. Wir sammelten Geschichten von Menschen, die auf YouTube „vom rechten Weg“ abkamen, insbesondere wenn dies ein Ergebnis ihres Empfehlungsalgorithmus war. Diese

Kampagne war unser erster Versuch herauszufinden, welche Art an Erfahrungen Nutzer:innen auf YouTube machen. Dabei gaben wir keine Vorgaben an, wovon die Geschichten handeln sollten; jede:r, der/die eine Geschichte beisteuerte, hatte sie selbst als „bedauerlich/unerfreulich“ bezeichnet. Seitdem haben wir uns bewusst dagegen entschieden zu definieren, was in die Kategorie „YouTube Regret“ fällt. Damit jede:r selbst bestimmen kann, was für sie/ihn als schlechte Erfahrung bei YouTube zählt.

Unser People-Powered-Ansatz stellt die gelebten Erfahrungen von Menschen in den Mittelpunkt, insbesondere die Erfahrungen von gefährdeten und/oder marginalisierten Menschen und Communitys, um die oft sehr legalistische und theoretische Diskussion darüber, was schädliche Inhalte im Internet sind, zu ergänzen. Unsere qualitative Analyse wurde in erster Linie von wissenschaftlichen Mitarbeitern der University of Exeter durchgeführt, die Videos anhand der YouTube-Community-Richtlinie, also der Regeln, die definieren, was auf der Plattform erlaubt ist, bewerteten. So bestimmten sie, ob ein Video auf der Plattform sein und/oder von YouTube empfohlen werden sollte oder nicht (näher beschrieben im Abschnitt „Analysemethoden“ dieses Berichts). Wenn Kommentare an die uns gemeldeten Videos angehängt waren, haben wir sie für unsere Analyse verwendet. Diese Methodik führt nicht zu objektiven Unterscheidungen – wir geben subjektive Eindrücke als Ausgangspunkt für weitere Gespräche und Studien wieder.

Während unsere Untersuchung zahlreiche Beispiele für Hassreden, widerlegte politische und wissenschaftliche Fehlinformationen und andere Kategorien von Inhalten aufgedeckt hat, die wahrscheinlich gegen die YouTube-Community-Richtlinien verstoßen (oder es tatsächlich tun), wurden auch viele Beispiele aufgedeckt, die ein komplizierteres Bild von Online-Risiken zeichnen. Viele der uns gemeldeten Videos fallen möglicherweise in die Kategorie dessen, was YouTube „grenzwertige Inhalte“ nennt – Videos, die die Grenzen der Community-Richtlinien „streifen“, ohne sie tatsächlich zu überschreiten. Da YouTube keine Transparenz darüber bietet, wie sie grenzwertige Inhalte definieren und klassifizieren, ist es unmöglich, diese Annahme zu überprüfen. Unsere Forschung legt auch nahe, dass bedauerliche oder schädliche Online-Erfahrungen oft das Ergebnis der Art und Weise sind, wie Inhalte im Laufe der Zeit auf jemanden ausgerichtet werden, was durch die Betrachtung einzelner Videos schwer zu verstehen oder zu beheben ist.

Die Empfehlungen, die wir in diesem Bericht aussprechen – mit denen wir Transparenz und Untersuchungen fordern und Kontrolle über Empfehlungsalgorithmen für Nutzer:innen –, sind von zentraler Bedeutung für die Menschen, die von diesen Systemen betroffen sind.



Ergebnisse

1. YouTube Regrets sind verstörend und unterscheiden sich voneinander

Die Zusammenfassung

„Jeden Tag kommen Millionen Menschen zu YouTube, um sich zu informieren, sich inspirieren oder unterhalten zu lassen“ – YouTube, 2020, „How YouTube Works“

- **Gemeldet wurde alles von Panikmache rund ums Thema COVID-19 über die Verbreitung politischer Fehlinformationen bis hin zu völlig unangemessenen „Kinder“-Cartoons.** Die am häufigsten auftretenden Kategorien an YouTube Regrets sind Fehlinformationen, brutale oder verstörende Inhalte, Hasstiraden und Spam/Betrug.

Die Story

Die Wahrnehmung der Internet-Plattformen durch die Öffentlichkeit hat sich in den letzten Jahren negativ verändert. Facebook und Twitter – die sich einst mit Lobgesängen von Nutzer:innen schmücken konnten – werden heute als Orte angesehen, an denen hasserfüllte Inhalte, Unhöflichkeit und Fehlinformationen vorherrschen.

YouTube hingegen hat es bisher geschafft, ein weithin positives Image zu behalten. Zumindest, wenn man es mit dem Image anderer Plattformen vergleicht. Allgemein gilt YouTube nach wie vor als ein Ort, an dem lustige Reaction-Videos und hilfreiche DIY-Inhalte florieren und als eine Community, in der wohlwollende Creator:innen erfolgreich sein können.

Und klar: YouTube kann wunderbar sein. Die Plattform bietet Millionen von Nutzer:innen täglich lehrreiche Inhalte und Unterhaltung. Aber genau wie die anderen sozialen Medien hat auch YouTube eine dunkle Seite – auch wenn diese vielleicht nicht so offensichtlich ist.

2019 haben wir uns auf den Weg gemacht, diese dunkle Seite zu inspizieren – um Videos besser zu verstehen, von denen mehr und mehr Nutzer:innen sprachen und die sie oft lieber nicht gesehen hätten. Wir nannten diese Art Inhalte „YouTube Regrets“ und riefen Menschen öffentlich dazu auf, uns von ihren Erfahrungen zu erzählen. Die Resonanz war alarmierend.

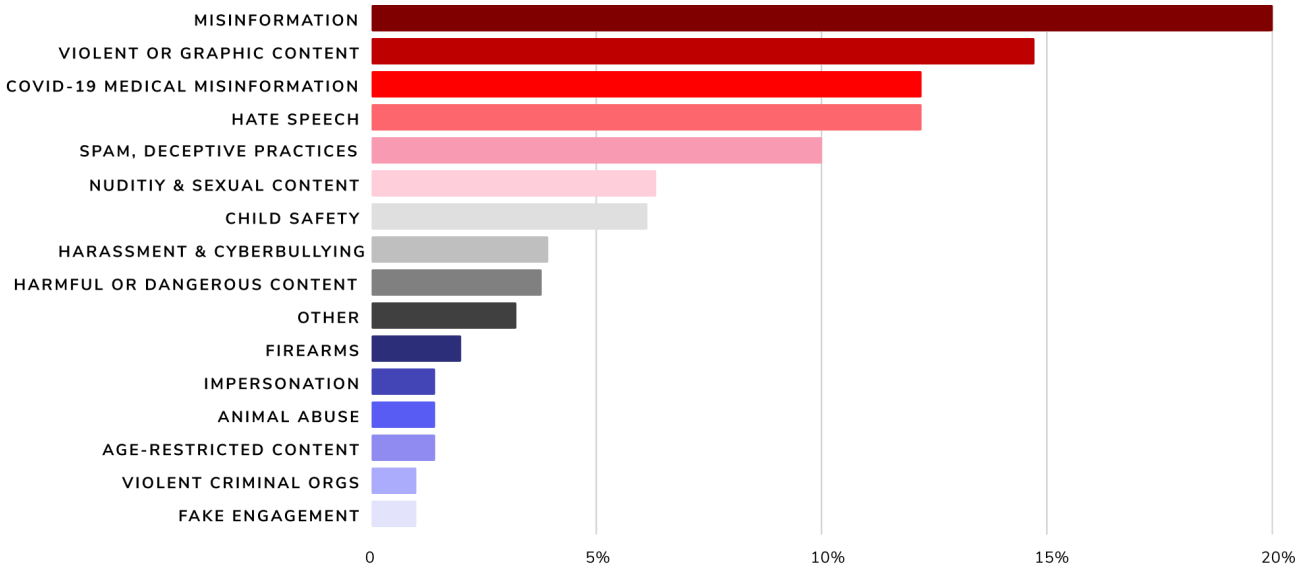
Neben lustigen und hilfreichen Videos gab es einen Berg von bizarren und sogar schädlichen Videos. Mozilla hörte von einer Person, die mehr und mehr Videos von Verkehrsunfällen vorgeschlagen bekam und „Videos sah, in denen Menschen den Unfall eindeutig nicht überlebten“. Ein 10-jähriges Mädchen, das Tanzvideos suchte, landete bei Videos, die extreme Diäten propagierten. „Sie schränkt jetzt ihr Essen und Trinken ein“, so ihre Eltern gegenüber Mozilla. Eine andere Person suchte nach bejahenden LGBTQ+-Inhalten, fand aber stattdessen hasserfüllte, Anti-LGBTQ+-Videos. „Ich kann mir nur vorstellen, wie schädlich dies für Menschen ist, die ihre Identität noch nicht gefunden haben“, schrieb er/sie an Mozilla.

Es gab viele weitere YouTube Regrets: pseudowissenschaftliche Videos, die Verschwörungstheorien über 9/11 beinhalten, misshandelte Tiere darstellen oder White Supremacy propagieren. Uns wurde klar, das YouTube ein echtes Problem hat. Deshalb entschieden wir uns dazu, das Thema genauer unter die Lupe zu nehmen, um die Erfahrungen, die echte Menschen auf der Plattform hatten und die Rolle, die YouTubes Algorithmus dabei spielt, zu untersuchen.

Die Daten

Wir schätzen, dass 12,2 % der gemeldeten Videos (95 % Konfidenzintervall von 10,4 bis 14,2 %) basierend auf den Community-Richtlinien von YouTube entweder „nicht auf

YouTube sein sollten“ oder „nicht proaktiv empfohlen werden sollten“. Die Kategorien, denen wir die Videos zuteilen können, sind in der folgenden Tabelle dargestellt. Obwohl die wissenschaftlichen Mitarbeiter:innen, die diese Videos klassifiziert haben, keine Experten für die Identifizierung und Klassifizierung von Fehlinformationen und die Einstufung illegaler Inhalte in Kategorien sind, basieren diese Klassifizierungen auf ihrem besten Urteilsvermögen und dem, was sie persönlich als beunruhigend und einer genaueren Untersuchung würdig empfanden. Wir stellen fest, dass die häufigsten Kategorien nach dieser Klassifizierung Fehlinformationen, gewalttätige oder grafische Inhalte und COVID-19-Fehlinformationen sind (die wir separat kategorisieren, da sie von besonderem Interesse sind), gefolgt von Hassreden und Spam/Betrug. Andere bemerkenswerte Kategorien sind Kindersicherheit, Belästigung und Cybermobbing und Tiermissbrauch. Es ist klar, dass YouTube Regrets aus einer Vielzahl von Gründen auftreten, aber dass Fehlinformationen ein dominierendes Problem sind. Wenn man die COVID-19-Fehlinformationen mit dem Rest der Fehlinformationen in einer einzigen Kategorie zusammenfasst, machen sie etwa ein Drittel aller kategorisierten Regrets aus.



Unten führen wir einige der deutlichsten Beispiele aus allen Kategorien auf. Unser Anhang enthält weitere Beispiele.

Einem/einer Freiwilligen wurde ein Video mit dem Titel „Omar Connected Harvester SEEN Exchanging $200 for General Election Ballot. 'We don't care illegal'", das eine unbegründete Behauptung über die amerikanische Aktivistin und Politikerin Ilhan Oman und Wahlbetrug bei den 2020 US-Wahlen verbreitet. In einem Kommentar schrieb

der/die Freiwillige, dass ihm/ihr andauernd rechtsextreme Kanäle empfohlen wurden, obwohl er/sie hauptsächlich Videos ansah, die von Überlebenskünsten in der Wildnis handeln. Das lässt darauf schließen, dass diese Empfehlungen Teil eines größeren Musters ist, das der/dem Freiwilligen politisch extreme Inhalte vorschlug.

Omar Connected Harvester SEEN Exchanging $200 for General Election Ballot."We don't care illegal."
542686 views - Sep 28, 2020

Jemand anderem wurde ein Video mit dem Titel „BILL GATES HIRED BLM “STUDENTS” TO COUNT BALLOTS IN BATTLEGROUND STATES“ vorgeschlagen, in dem die unbegründete und anklagende Behauptung verbreitet wird, dass der Gründer von Microsoft junge Menschen aus ethnischen Minderheiten dafür angestellt habe, Wahlbetrug zu begehen.

BILL GATES HIRED BLM “STUDENTS” TO COUNT BALLOTS IN BATTLEGROUND STATES
14 views - Nov 21, 2020

Jemand meldete ein Video namens „The Elites Who Control You“, in dem eine als Politiker oder „elitäre Person“ verkleidete Person angibt, das sie COVID-19 nutzen und Lügen verbreiten würde, um damit Angst zu verbreiten und so die Massen zu lenken.

The Elites Who Control You
743641 views - Dec 1, 2020

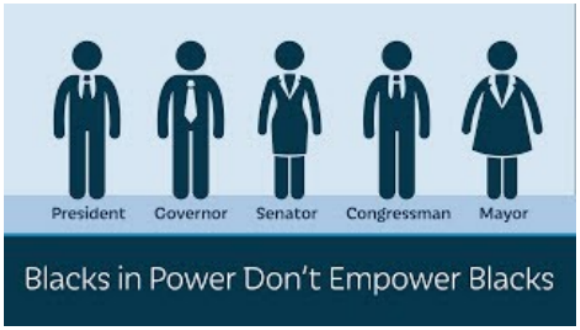
Jemand meldete ein Video namens „7 jokes ending in tragedy“. Die in diesem Video gezeigten Geschichten sind sehr verstörend und im Vorschaubild des Videos ist eine große Menge an Blut zu sehen.

7 SCHERZI finiti in TRAGEDIA
2193169 views - 6 apr 2018

Jemand meldete ein animiertes Video mit dem Titel „Woody's Got Wood“, einer sexualisierten Parodie auf den Kinderfilm „Toy Story“.

Woody's Got Wood Animated.mp4
22251 views - Sep 17, 2018

Jemand meldete ein empfohlenes Video mit dem Namen „Blacks in Power Don't Empower Blacks“, in dem rassistische und beleidigende Sprache und Ideen geäußert werden.



Blacks in Power Don't Empower Blacks
2365794 views - Mar 26, 2018

Jemandem wurde ein Video mit dem Titel „El Arca - It’s Literally Furry Noah’s Arc“ empfohlen, was sie als bizarr einstuften. Das Video legt nahe, dass der Kinderfilm, der darin besprochen wird, sexuelle und erwachsene Inhalte enthält und erläutert diese gründlich.

El Arca - It's Literally Furry Noah's Arc
119145 views - 22 Apr 2021

Ein Video mit dem Titel „‘Biggest fraud’ in US history—up to 300,000 fake people voted in Arizona election: expert | NTD“ wurde einem/einer freiwilligen Helfer:in vorgeschlagen.

'Biggest fraud' in US history—up to 300,000 fake people voted in Arizona election: expert | NTD
965541 views - 1 Dec 2020

Jemand meldete eine Video mit dem Titel „[Documentary 2020] [Adrenochrome — The Darkest Drug of Them All!]“, das von einer durch QAnon (einer mutmaßlich amerikanischen Person oder Gruppe, die seit 2017 tätig ist) verbreiteten Verschwörungstheorie darüber handelt, dass Kinderblut „geerntet“ wird.

[Documentary 2020] [Adrenochrome - The Darkest Drug of Them All!]
108279 views - Premiered Mar 23, 2020

2. Der Algorithmus ist das Problem

Die Zusammenfassung

„Wenn Empfehlungen tun, was sie sollen, können sie Nutzer:innen bei der Entdeckung eines neuen Lieblingslieds, neuer toller Creator:innen oder leckerer Paella-Rezepte helfen. Deshalb aktualisieren wir unser System für Empfehlungen regelmäßig. Wir möchten Nutzer:innen Videos empfehlen, die sie auch sehen möchten.“ –YouTube, 2019, „Continuing our work to improve recommendations on YouTube“

- **Ungefähr 9 % der Regrets, die über YouTubes Algorithmus empfohlen wurden, darunter einige, die gegen die Community-Richtlinien der Plattform verstoßen, wurden seitdem von YouTube entfernt.** Insgesamt hatten diese empfohlenen und später entfernten Regrets zu dem Zeitpunkt, an dem sie gemeldet wurden, 160 Millionen Ansichten.
- **Empfehlungen sind unverhältnismäßig oft für YouTube Regrets verantwortlich.** 71 % aller gemeldeten YouTube Regrets beziehen sich auf Videos, die unseren freiwilligen Teilnehmer:innen empfohlen wurden. Außerdem war die Wahrscheinlichkeit, dass die Wiedergabe empfohlener Videos bedauert werden würde, 40 % höher als bei Videos, nach denen Nutzer:innen gesucht hatten.
- **Dabei erzielen solche Videos auf YouTube gute Ergebnisse:** Gemeldete Videos erreichen bis zu 70 % mehr Ansichten pro Tag als andere Videos, die freiwillige Helfer:innen ansahen.
- **Bedauerliche Empfehlungen sind oft unprovoziert.** Aus den Daten, die wir von freiwilligen Helfer:innen erhalten haben, geht hervor, dass in 43,3 % aller Fälle, in denen die Ansicht von Inhalten im Nachhinein bereut wurde, die Empfehlung absolut nichts mit den Inhalten/Videos zu tun hatte, die der/die Nutzer:in vorher angeschaut hatte.

Die Story

Eine Recherche der HuffPost berichtete 2020 die Geschichte eines 11-jährigen Mädchens namens Allie. Allie filmte unwissentlich Sketche, in denen sie sexuelle Fantasien von Pädophilen nachstellte, die ihren YouTube-Kanal gefunden hatten und sie zum Beispiel baten, vor der Kamera so zu tun, als falle sie in Ohnmacht. Noch schlimmer ist es, dass die Untersuchung der HuffPost ergab, dass der YouTube-Algorithmus solchen Tätern dabei *half*, Kanäle wie den von Allie zu finden. Diese verstörende Untersuchung deckte die negative Seite von algorithmusbedingten Empfehlungen auf und zeigte, was passieren kann, wenn scheinbar „harmlose“ Videos, die nicht gegen YouTubes Community-Richtlinien verstoßen – wie etwa Videos, die badende oder turnende Kinder oder Kinder zeigen, die ihre Beine spreizen –, algorithmisch eingeordnet und Menschen gezeigt werden, die von solchen Inhalten besonders angesprochen werden.

Allies Geschichte ist ein verstörendes Beispiel aus dem echten Leben, das zeigt, wie YouTubes automatische Empfehlungen dazu beitragen können, Videos „viral gehen“ zu lassen, ohne dabei den Kontext zu bedenken, in dem diese Videos angesehen werden. Da YouTube ein offenes System für die Empfehlung von Inhalten ist, also nutzer:innengenerierte Videos empfiehlt, ohne dass diese vorher überprüft werden, können die vom Algorithmus getroffenen Entscheidungen größere (und riskantere) Konsequenzen haben als die auf Plattformen wie Netflix, die ausschließlich von Menschen freigegebene Inhalte empfehlen.

Viele Plattformen, auf denen nutzer:innengenerierte Inhalte automatisch empfohlen werden, verlassen sich dabei auf Algorithmen, um Inhalte zu identifizieren und zu moderieren. Allerdings sind diese Algorithmen sehr begrenzt und müssen neben Algorithmen funktionieren, die Nutzer:innen basierend auf Daten wie der Dauer, für die ein Video angesehen wird, „interessante“ Inhalte vorschlagen. Zusammengenommen bedeutet das, dass Algorithmen Vorhersagen darüber abwägen müssen, wie wahrscheinlich ein Video von Nutzer:innen als „interessant“ empfunden wird oder wie wahrscheinlich ein Video gegen YouTubes Community-Richtlinien verstößt. Das ist eine große Verantwortung und eine, die kommerzielle Interessen gegen das Wohlbefinden der Nutzer:innen ausspielen kann.

Unsere Recherchen ergaben mehrere Fälle, in denen der YouTube-Algorithmus diese doppelte Verantwortung schlecht auszugleichen schien. Wir konnten mehrere Fälle identifizieren, in denen der YouTube-Algorithmus Videos empfohlen hat, die tatsächlich gegen die eigenen Community-Richtlinien und andere Inhaltsrichtlinien verstießen und später von der Plattform entfernt wurden, nachdem sie bereits Millionen von Aufrufen erreicht hatten. Unsere Untersuchung ergab auch, dass die Empfehlungen des YouTube-Algorithmus im Vergleich mit etwa einer gezielten Suche nach Videos oder anderen Wegen, auf die unsere freiwilligen Helfer:innen auf Inhalte aufmerksam wurden, unverhältnismäßig oft für die schlechten Erfahrungen verantwortlich waren. Wir fanden auch heraus, dass YouTube Regrets, von denen unsere Freiwilligen berichteten, im Vergleich zu anderen Videos auf der Plattform schnell eine beträchtliche Anzahl an Ansichten gewonnen hatten. Diese Beispiele werfen ernsthafte Fragen darüber auf, wie YouTube die verschiedenen Entscheidungen priorisiert, die sein Algorithmus treffen muss, und welche Kompromisse dabei eingegangen werden.

YouTubes Algorithmus hält die Fäden in der Hand. Welche Inhalte er vorschlägt, warum und wie diese Entscheidungen getroffen werden, ist wirklich wichtig – insbesondere in Situationen wie der von Allie. Während einer Anhörung im US-Senat im April 2021 hielt Senator Chris Coons (D-DE) YouTube dazu an, sich zur Freigabe von Informationen darüber, wie oft die Plattform ein beliebiges Video *empfiehlt* und nicht nur darüber, wie häufig ein solches Video *angesehen* wird, zu verpflichten. Diese Information ist wichtig, um die Bedeutung des YouTube-Algorithmus für virale Videos nachzuvollziehen und könnte dabei helfen zu verstehen, wie beispielsweise Allies Videos von YouTube vorgeschlagen wurden. YouTubes Vertreter:in wollte sich während der Anhörung nicht zu der Bereitstellung dieser Informationen verpflichten.

Die Daten

Mehrere der von uns gemeldeten Regrets wurden seitdem aus YouTube entfernt. YouTube nutzt eine Mischung aus Menschen und maschinellem Lernen, um Inhalte zu identifizieren, die gegen die Richtlinien verstoßen, und diese von der Plattform zu löschen. Allerdings machen diese Systeme oft Fehler. Im Zusammenspiel mit automatisch generierten Empfehlungen schlägt YouTube dann häufig Inhalte vor, die gegen die eigenen Vorschriften verstoßen. Bis zum 1. Juni 2021 wurden insgesamt 189 der an unsere freiwilligen Helfer:innen empfohlenen und an uns gemeldeten Videos gelöscht.

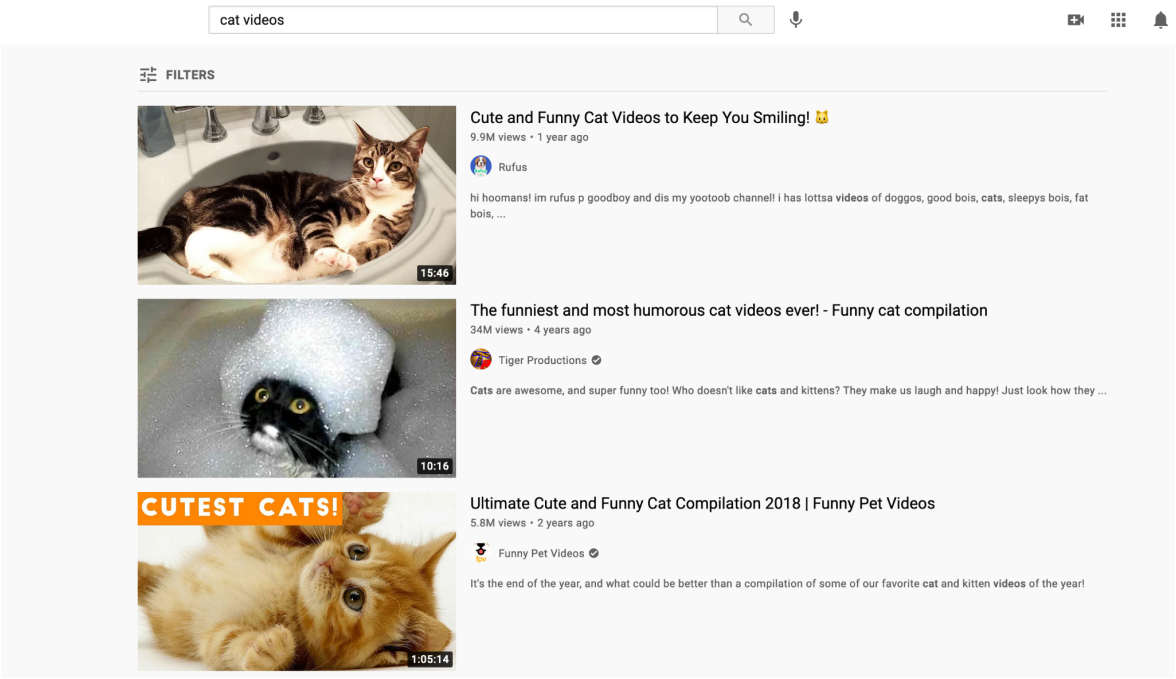
In einigen Fällen gibt YouTube Gründe für die Löschung eines Videos an (z. B. die Verletzung von Community-Richtlinien, Urheberrechtsverletzung, Bestimmungen zu Verhetzung oder dass ein Video von dem/der Nutzer:in gelöscht oder auf privat gestellt wurde). Aber in fast 40 % der von uns analysierten Fälle wurden die Videos einfach als „Video nicht verfügbar“ gelabelt ohne nähere Angaben dazu, warum das Video von der Plattform entfernt wurde.

Diese Videos verzeichneten über die durchschnittlich 5 Monate, in denen sie sich auf der Plattform befanden, zusammengenommen 160 Millionen Aufrufe, als sie gemeldet wurden, also durchschnittlich 760.000 Aufrufe pro Video. Wie viele dieser Aufrufe das Ergebnis von Empfehlungen durch den YouTube Algorithmus waren, lässt sich nicht feststellen, da YouTube diese Daten nicht veröffentlicht. Wir wissen aber, dass sie mindestens einmal empfohlen wurden (an unsere:n freiwillige Helfer:in, der/die sie uns meldete).

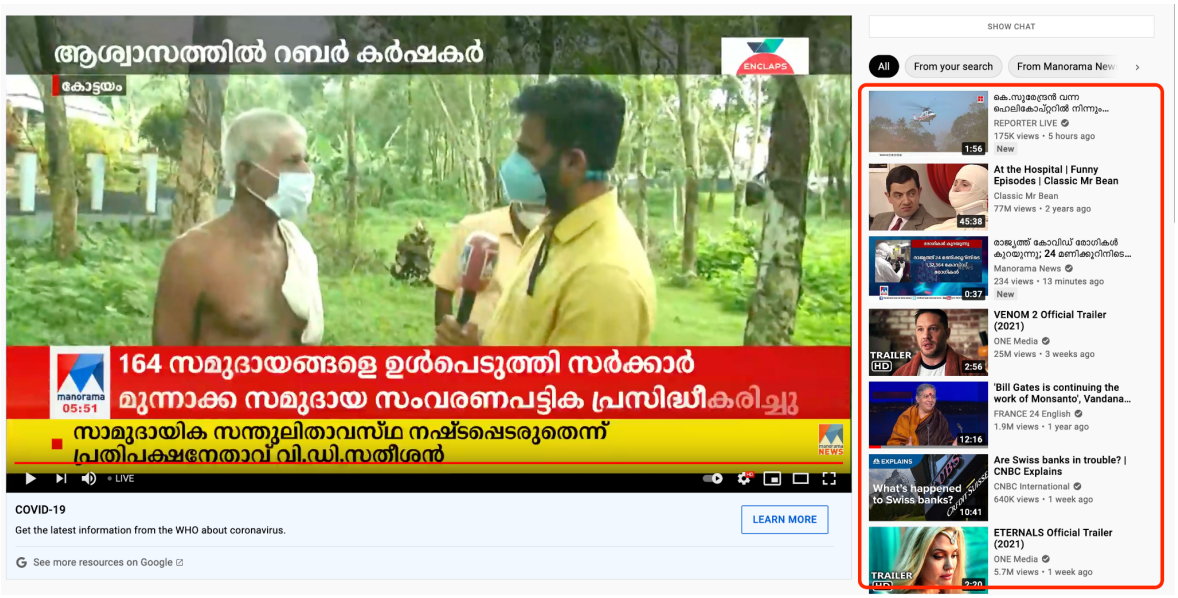
Mit unseren Daten können wir berechnen, welchen Anteil der insgesamt angesehenen Videos unsere freiwilligen Helfer:innen als Regret melden (oder als Video, dass sie sich lieber nicht angeschaut hätten). Dieser Anteil kann in verschiedene Faktoren unterteilt werden, beispielsweise wie die/der Freiwillige das Video gefunden hat (z. B. Suche, Empfehlungen, Link). Dabei haben wir unsere Analyse auf zwei Hauptpunkte konzentriert:

- Suchen sind Fälle, in denen der/die Freiwillige eine Suchanfrage eingibt und dann eins der Videos ansieht, die YouTube als Antwort auf die Anfrage anzeigt.
- Empfehlungen sind Fälle, in denen YouTube einer/einem Freiwilligen proaktiv Inhalte vorschlägt.

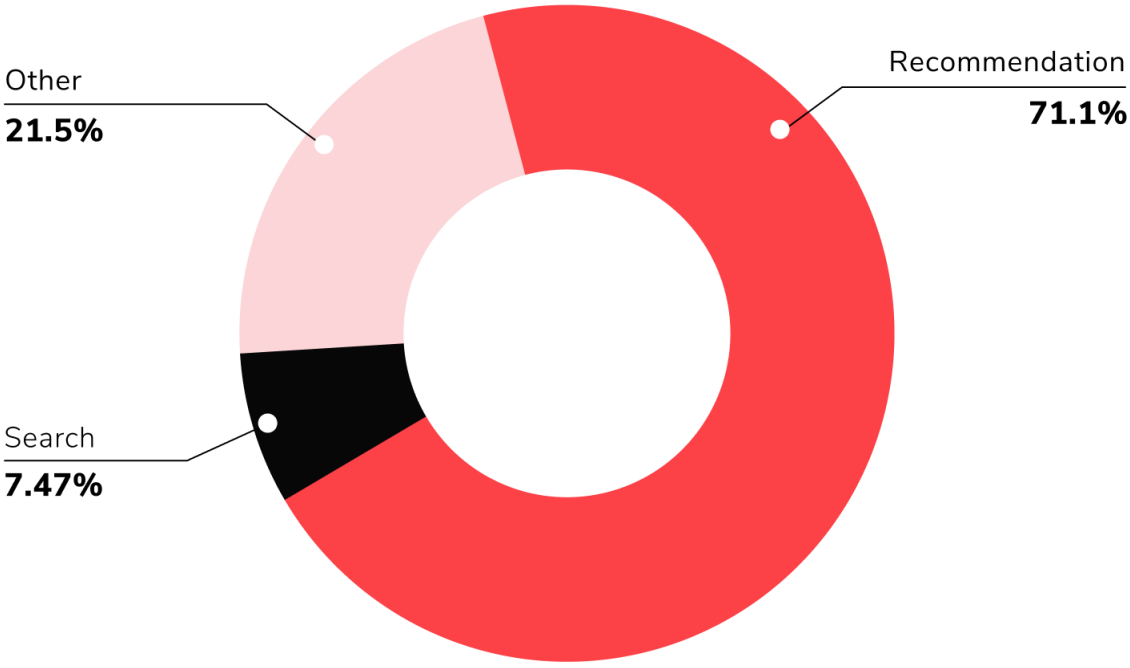
Suche



Empfehlung



Durch den Vergleich der Anteile an Regrets, die nach dem Aufruf eines empfohlenen Videos entstehen, mit denen, die bei Suchen entstehen, wird deutlich: Empfehlungen sind unverhältnismäßig oft für YouTube Regrets verantwortlich. Bei Suchen werden 9,6 von 10.000 angesehenen Videos als Regret gemeldet. Bei Empfehlungen liegt der Anteil an Regrets bei 13,9 von 10.000 angesehenen Videos und ist damit 40 % höher. Von sämtlichen uns gemeldeten Regrets sind außerdem rund 71 % Empfehlungen.



Als Anhaltspunkt: Die von YouTube veröffentlichte „Violative View Rate“, die die Anzeigequote für Videos darstellt, die gegen die YouTube-Richtlinien verstoßen, liegt bei etwa 17 Videos von 10.000. Die Tatsache, dass unsere Quote (13,9 von 10.000 für Empfehlungen) nicht über den von YouTube selbst veröffentlichten Ergebnissen liegt, unterstützt die Gültigkeit unseres Konzepts der YouTube Regrets.

Gemeldete Videos generieren außerdem schneller mehr Aufrufe als andere Videos. YouTube Regrets hatten bis zum Tag, an dem sie gemeldet wurden, durchschnittlich 5.794 Aufrufe pro Tag, an dem sie auf der Plattform waren. Das ist 70 % mehr als andere von unseren Freiwilligen angesehene Videos (die angesehen wurden, bevor das empfohlene Video gemeldet wurde). Diese hatten durchschnittlich nur 3.312 tägliche Aufrufe.

Bei der Analyse dieser nacheinander angesehenen Videos fiel uns auf, dass einige Empfehlungen keinen Bezug zu dem hatten, was der/die Freiwillige vorher angesehen hatte. Um diesen Trend genauer zu untersuchen, baten wir wissenschaftliche Mitarbeiter:innen darum, die Videos zu kategorisieren, die bis zum gemeldeten Regret angesehen wurden. So wollten wir feststellen, ob die Empfehlungen einen Bezug hatten oder nicht. Unter Empfehlungen, für die uns Daten über die vorherigen Videos vorliegen, die der/die Freiwillige ansah, hatten die Regrets in 43,3 % der Fälle keinen Bezug zu vorher angesehenen Videos.

Ein:e Freiwillige:r sah sich ein Musikvideo von Art Garfunkel an und bekam anschließend die Empfehlung, sich ein Video mit dem Titel „Trump Debate Moderator EXPOSED as having Deep Democrat Ties, Media Bias Reaches BREAKING Point“ anzusehen. Kommentare, die Freiwillige mit den Videovorschlägen einreichten, zeigten, dass sie von den vorgeschlagenen Videos ohne Bezug auf die vorher angesehenen Videos erschöpft und verärgert waren, insbesondere wenn diese Videos ihren Glaubenssätzen widersprachen oder einen Standpunkt vertraten, dem sie nicht zustimmten.

The video being reported

Trump Debate Moderator EXPOSED as having Deep Democrat Ties, Media Bias Reaches BREAKING Point

166310 views - Oct 18, 2020

Video history for this session

Art Garfunkel and his son cover The Everly Brothers live in Napa, May 12, 2019 (4K)

Nach der Wiedergabe eines Mozilla-Videos mit dem Titel „Memes, Misinfo, and the Election - October 12, 2020“ wurde einem Teilnehmer ein Video namens „Global Warming: Fact or Fiction? Featuring Physicists Willie Soon and Elliott D. Bloom“ empfohlen. In einem Kommentar schrieb der/die Freiwillige, dass er/sie davon überrascht war, ein Video vorgeschlagen zu bekommen, in dem der Klimawandel abgestritten wird, nachdem er/sie ein Mozilla-Video angesehen hatte.

The video being reported

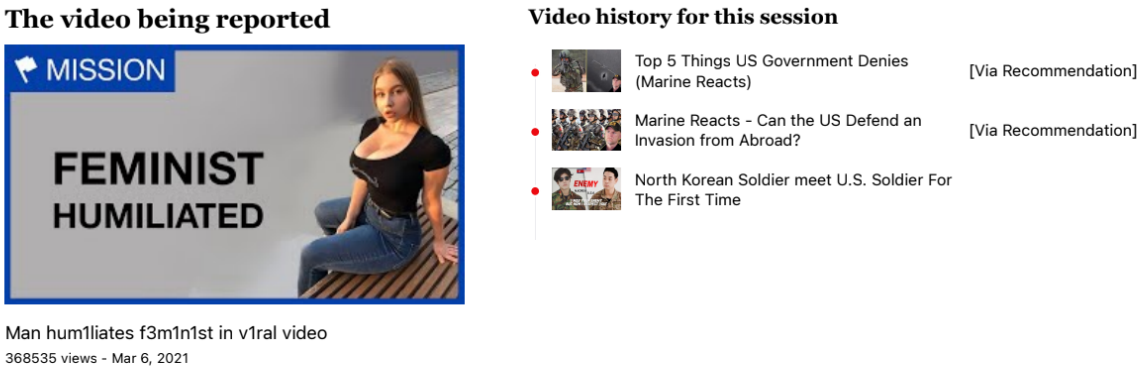
Global Warming: Fact or Fiction? Featuring Physicists Willie Soon and Elliott D. Bloom

592557 views - Aug 16, 2019

Video history for this session

Memes, Misinfo, and the Election - October 12, 2020

Nachdem sich ein Teilnehmer Videos über das US-Militär angesehen hatte, wurde ihm ein Video mit dem Titel „Man humiliates feminist in viral video“ vorgeschlagen, das misogyn, sexistisch und Frauen gegenüber diskriminierend ist.



3. Nicht-englischsprechende Nutzer:innen sind am stärksten betroffen

Die Zusammenfassung

„YouTubes Community-Richtlinien werden weltweit durchgehend durchgesetzt, egal wo die Inhalte hochgeladen werden.“ – YouTube, 2021, „YouTube Community Guidelines Enforcement Transparency Report“

- **Nicht-englischsprechende Nutzer:innen sind am stärksten betroffen.** Die Anzahl an Videos, die sich Nutzer:innen im Nachhinein lieber nicht angesehen hätten, liegt in Ländern, in denen Englisch nicht die Primärsprache ist, um 60 % höher.
- **Besonders viele Meldungen gibt es in nicht-englischsprachigen Ländern zu pandemiebezogenen Inhalten.** Von den YouTube Regrets, die sich nach unseren Erkenntnissen nicht auf YouTube befinden sollten bzw. nicht von YouTube empfohlen werden sollten, waren laut unseren Untersuchungen nur 14 % der englischsprachigen Videos pandemiebezogen. Unter den nicht englischsprachigen Videos liegt die Quote bei 36 %.

Die Story

Im Jahr 2017 waren etwa 700.000 Menschen aus der Rohingya-Gemeinschaft, einer ethnischen muslimischen Minderheit in Myanmar, gezwungen, ins benachbarte Bangladesch zu fliehen, um Tötungen, Massenvergewaltigungen und der Verbrennung ihrer Dörfer zu entgehen. Monate später veröffentlichte die unabhängige internationale Untersuchungskommission der UNO für Myanmar einen Bericht, in dem hervorgehoben wurde, wie Facebook den Gräueltaten des Militärs in Myanmar im Rakhine-Staat, der Heimat der Rohingya und anderer ethnischer Minderheiten, eine Plattform bat und sie so begünstigte. Eine Untersuchung von Reuters fand „toxische Beiträge, die die Rohingya oder andere Muslime als Hunde, Maden und Vergewaltiger bezeichneten, vorschlugen, sie an Schweine zu verfüttern, und darauf drängten, sie zu erschießen oder auszurotten“. Die Hasstiraden, die auf Facebook gediehen und sich ausbreiteten, wurden allgemein als ein wesentlicher Vorläufer des daraus resultierenden Völkermordes genannt. Wie Reuters feststellte, waren fast alle diese Beiträge in der lokalen Primärsprache, Burmesisch, verfasst. Und fast alle von ihnen verletzten die Richtlinien von Facebook.

Um es auf den Punkt zu bringen: Die Richtlinien von Plattformen und die Durchsetzung dieser Richtlinien sind zwei verschiedene Paar Schuhe. Und leider werden die Richtlinien der Plattformen in verschiedenen Teilen der Welt sehr unterschiedlich durchgesetzt. Ein Grund dafür ist, dass Algorithmen, die zur Erkennung von Richtlinienverletzungen und zur Empfehlung von Videos verwendet werden, auf sprachspezifischen maschinellen Lernmodellen basieren. Das bedeutet, dass Unternehmen ihre Algorithmen mit Daten aus verschiedenen Sprachen und Länderkontexten trainieren müssen. Viele Plattformen priorisieren jedoch das Training auf englischsprachigen Daten, weshalb sie in diesen Kontexten besser abschneiden.

YouTube hat zwar keine Daten dazu veröffentlicht, räumt aber ein, dass seine Empfehlungssysteme zur Verbreitung grenzwertiger Inhalte beitragen, also Inhalten, die die Grenzen der Community-Richtlinien streifen, ohne sie tatsächlich zu überschreiten. Als das Unternehmen Änderungen an den Richtlinien ankündigte, um dieses Problem anzugehen, konzentrierte es sich zunächst auf die Vereinigten Staaten und andere englischsprachige Länder. Mehr als zwei Jahre später gab das Unternehmen bekannt, dass es diese Änderungen in allen Märkten, in denen es tätig ist,

eingeführt hat. Unseres Wissens nach hat YouTube jedoch keinerlei Metriken über den Erfolg dieser Bemühungen außerhalb der Vereinigten Staaten veröffentlicht.

Unsere Untersuchung ergab, dass YouTube Regrets in nicht-englischsprachigen Ländern um 60 % häufiger sind. Außerdem fanden wir heraus, dass bei den Videos, die nach Meinung unserer wissenschaftlichen Mitarbeiter:innen nicht auf YouTube zu sehen sein oder von YouTube empfohlen werden sollten, Videos im Zusammenhang mit der Pandemie in nicht-englischsprachigen Ländern häufiger vorkamen.

Wenn Menschen weltweit auf YouTube nach wichtigen Informationen suchen, wie beispielsweise Informationen über eine globale Pandemie, können diese Diskrepanzen desaströs sein. Ein kürzlich veröffentlichter Bericht des EU Disinfo Lab deckte auf, wie ein französischsprachiges Video mit dem Titel „Hold Up“ (ähnlich dem englischsprachigen Video „Plandemic“), in dem es um COVID-19-Verschwörungen geht und das gegen YouTubes Richtlinien zur medizinischen Fehlinformation verstößt, immer noch auf der Plattform verfügbar ist. Und das sechs Monate nach seiner Veröffentlichung und als es bereits Millionen Aufrufe generiert hatte. Das stimmt mit Untersuchungen überein, die von der Election Integrity Partnership durchgeführt wurden und die belegen, dass nicht-englischsprachige Fehlinformationen mit Bezug auf die amerikanischen Wahlen 2020 auf Twitter und YouTube weder gekennzeichnet waren noch durch andere Maßnahmen gegen sie vorgegangen wurde. Außerdem war unklar, auf welchen Sprachen YouTubes Richtlinien durchgesetzt wurden.

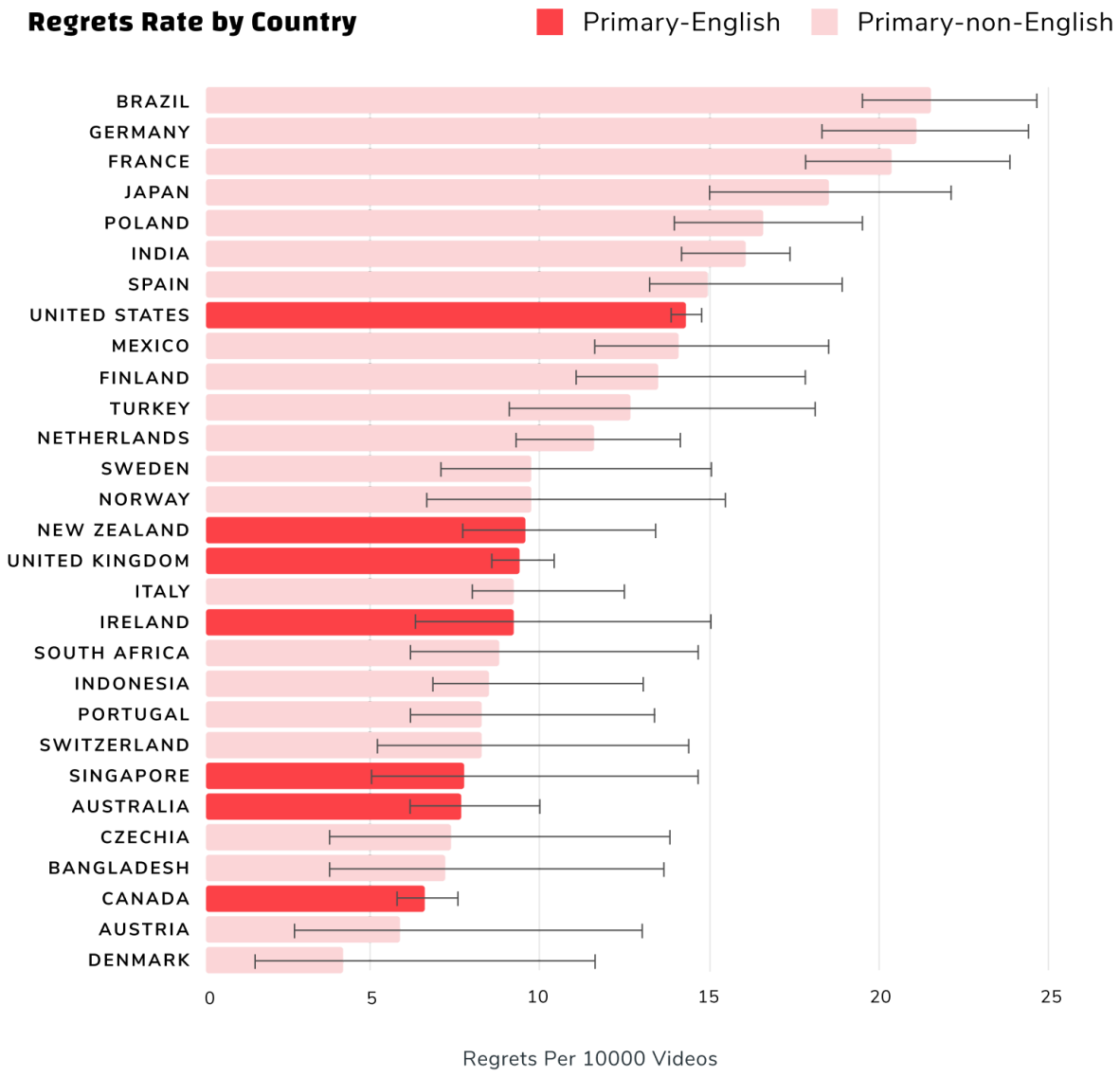
Es sollte nicht von der Sprache einer Person abhängen, ob er/sie online geschützt wird oder nicht. Plattformen wie YouTube sind dafür verantwortlich alle Nutzer:innen ihrer Plattform zu schützen, nicht nur solche, die in englischsprachigen Ländern leben und Englisch sprechen. Die systematisch nicht-vorhandene Transparenz und Aufmerksamkeit auf dieses Problem verstärkt die Wichtigkeit von Tools wie dem RegretsReporter, mit dessen Hilfe Nutzer:innen selbst überwachen können, wie und ob Plattformen ihre eigenen Richtlinien konsequent durchsetzen oder nicht.

Die Daten

Wir haben die Regret-Anteile nach Land mithilfe der GeoIP-Suche berechnet, die uns verrät, aus welchem Land unsere Freiwilligen auf YouTube zugreifen. Wir fanden heraus, dass die höchsten Regret-Quoten – über 20 Regrets pro 10.000 angeschaute Videos – in Brasilien, Deutschland und Frankreich zu finden sind. Tatsächlich haben die

sieben führenden Länder alle eine nicht-englische Primärsprache. Die USA sind das einzige englischsprachige Land in den Top 14 der höchsten Regret-Quoten. Unter den 14 Ländern mit den niedrigsten Regret-Quoten befinden sich fünf primär englischsprachige Länder.

Es ist eindeutig, dass die Regret-Quoten in Ländern, deren Primärsprache Englisch ist, niedriger ist. Unter den Ländern, in denen Englisch Primärsprache ist, liegt die Quote bei 11,0 Regrets pro 10.000 wiedergegebenen Videos (Konfidenzintervall von 95 % liegt bei 10,4 bis 11,7). In Ländern, in denen Englisch nicht Primärsprache ist, ist die Quote 17,5 Regrets pro 10.000 wiedergegebenen Videos (Konfidenzintervall von 95 % liegt bei 16,8 bis 18,3). Wir sehen also eine klare statistisch signifikante Differenz: In Ländern, in denen Englisch nicht die Primärsprache ist, liegt die Quote an Regrets um 60 % höher.



Die Fehlerbalken zeigen ein 95%iges Konfidenzintervall für die Regret-Quote. Aufgrund der begrenzten Daten in vielen Ländern sind diese Konfidenzintervalle in einigen Fällen sehr breit, aber die Ergebnisse sind alle statistisch signifikant. Beachten Sie, dass wir nur die Länder zeigen, für die wir ausreichend Daten haben, um genauere Schätzungen, gemessen an der Konfidenzintervallbreite, vorzunehmen.

Wir bestimmten die Primärsprache eines Landes mithilfe des CIA World Factbook; dabei wählen wir die Sprache als Primärsprache eines Landes, die im Verhältnis zur Einwohnerzahl von den meisten Menschen wenn möglich gesprochen wird oder am weitesten verbreitet ist. Unsere Kategorisierung umfasst die Länder, die den Großteil

unserer Daten liefern (verantwortlich für 94 % der Regret-Berichte), entweder als englische oder nicht-englische Primärsprache.

Pandemie-bezogene Inhalte waren ein bemerkenswertes Segment unserer Regret-Berichte und waren besonders in nicht-englischen Videos weit verbreitet. Wir haben die Videos analysiert, von denen unsere wissenschaftlichen Mitarbeiter:innen entschieden haben, dass sie entweder nicht auf YT sein oder nicht empfohlen werden sollten, und haben herausgefunden, dass unter den Videos in englischer Sprache nur 14 % einen Pandemiebezug haben. Andererseits liegt die Quote bei als Regret gemeldeten Videos, die nicht auf Englisch sind, bei 36 % – mehr als ein Drittel dieser Regrets in anderen Sprachen als Englisch haben mit der Pandemie zu tun.

In dem portugiesischen Livestream-Video namens „Corrupt biden, judicialization of the vaccine and enough of hysteria!," verteidigt die Person Verschwörungstheorien über das Coronavirus.

The video being reported

Biden corrupto, judicialização da vacina e chega de histeria!

6 views - Stream iniciado há 57 minutos

Dieses slowakische Video wurde uns gemeldet. Es handelt davon, dass die Pandemie nicht wirklich ist, sondern es sich dabei um eine Falschmeldung handelt. Als uns das Video gemeldet wurde, hatte es bereits über 300.000 Aufrufe.

The video being reported

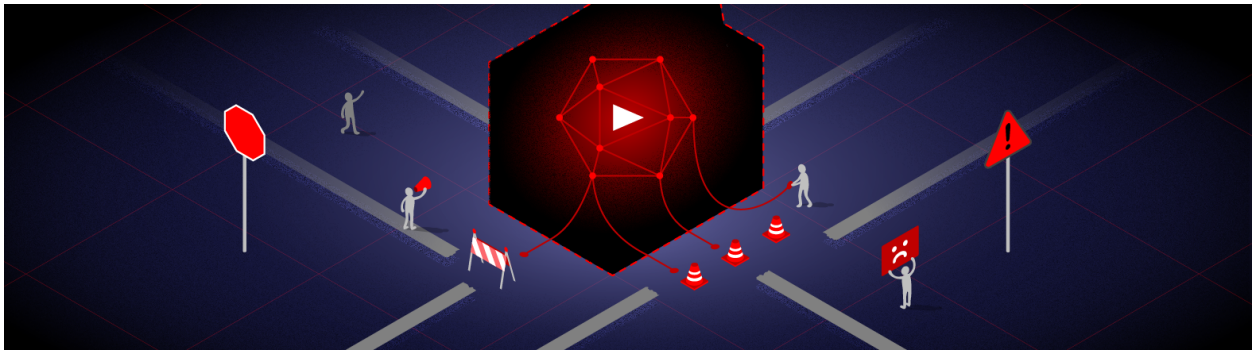
Celoplošné testovanie: fiasko a konflikt od prvej chvíle
300385 views - 19 Oct 2020

In dieser griechischen Aufnahme eines Podcasts/einer Vorlesung mit Thema Pandemie suggeriert der/die Sprecher:in, dass die Pandemie ein praktisches Mittel für die griechische Regierung sei, ihre Agenda weiterzuverfolgen.

The video being reported

ΧΡΥΣΗ ΑΥΓΗ ΚΑΣΙΔΙΑΡΗΣ ΚΟΡΩΝΟΙΟΣ ΝΤΟΚΟΥΜΕΝΤΑ & ΑΛΗΘΕΙΑ στην ΕΠΙΔΗΜΙΑ ΑΠΑΤΗ φοβου των ΚΑΝΑΛΙΩΝ 2η
1916 views - 5 Απρ 2020

Es gibt für diese Ergebnisse viele mögliche Erklärungen, darunter auch kulturelle Unterschiede in Bezug auf Meldungen oder Internetnutzung. Allerdings weisen unsere Daten darauf hin, dass YouTube in nicht-englischsprachigen Märkten ein größeres Problem mit Regrets hat.



Empfehlungen

Unsere Untersuchung legt nahe, dass die Unternehmensrichtlinien und -praktiken von YouTube, einschließlich der Gestaltung und des Betriebs ihrer Empfehlungsalgorithmen, zumindest teilweise für die bedauerlichen Erfahrungen verantwortlich sind, die unsere Freiwilligen auf der Plattform gemacht haben. Wir glauben, dass unsere Untersuchung nur die Spitze des Eisbergs aufgedeckt hat und dass jedes dieser Ergebnisse eine weitere Untersuchung verdient und notwendig macht. Wir sind uns auch darüber im Klaren, dass ohne ein Eingreifen, das eine genauere Überprüfung der Algorithmen von YouTube ermöglicht, diese Probleme weiterhin unkontrolliert bleiben und die Folgen für unsere Communitys zunehmen werden. Trotz des Fortschritts, den YouTube bei diesen Themen angeblich gemacht hat, ist es für Forscher:innen immer noch fast unmöglich, diese Behauptungen zu überprüfen oder die Empfehlungsalgorithmen von YouTube zu untersuchen.

Unsere Empfehlungen an YouTube und andere Plattformen

1. **Unabhängige Überprüfungen von Empfehlungsalgorithmen müssen möglich gemacht werden.**

   Für eine echte Rechenschaftspflicht ist es unerlässlich, dass Aufsichtsbehörden und Forscher:innen, die im öffentlichen Interesse tätig sind, bei Plattformen wie YouTube „hinter die Kulissen blicken“ können, um besser verstehen und bewerten zu können, wie das Design und die operative Praxis, die mit ihren Empfehlungssystemen verbunden sind, zu Online-Risiken beitragen können.

   Zu diesem Zweck sollten YouTube und andere Plattformen eine gemeinschaftliche Anstrengung unternehmen, um die Bemühungen von

Forscher:innen zu erleichtern, die mit ihren Empfehlungssystemen verbundenen Risiken und Schäden zu untersuchen. Entscheidend ist, dass diese Datenzugriffsregelungen robust sind und weit über das hinausgehen, was von Plattformen wie YouTube heute angeboten wird, denn wie unser Bericht gezeigt hat, gibt es einfach zu viele Lücken.

Anstelle von bruchstückhaften und oberflächlichen Einblicken sollten Forscher:innen – unter der Bedingung, dass sie sich an Datenschutz- und Sicherheitsprotokolle halten – befugt sein, Tests mit Empfehlungssystemen durchzuführen, Zugang zu Informationen über die von Empfehlungssystemen optimierten Metriken zu erhalten, auf den Quellcode des Empfehlungssystems zuzugreifen, auf Trainingsdaten zuzugreifen, die zum Trainieren von maschinellen Lernmodellen verwendet werden, und die Dokumentation in Bezug auf Vertrauens- und Sicherheitspraktiken einzusehen.

Diese Art der Transparenz würde uns dabei helfen, versteckte Risiken zu identifizieren und kommerzielle und regulatorische Lösungen zu entwickeln, um diese anzugehen.

2. **Plattformen müssen Informationen darüber offenlegen, wie ihre Systeme für Empfehlungen funktionieren, und transparente Berichte erstellen, die einen ausreichenden Einblick in Problemgebiete und den Fortschritt auf diesen Gebieten gewähren.**

   Empfehlungssysteme sind der wichtigste Mechanismus zur Bereitstellung von Inhalten für Plattformen wie YouTube und ein entscheidender Ort für Vertrauens- und Sicherheitsinterventionen. In Anerkennung der Bedeutung von Empfehlungssystemen bei der Verbreitung schädlicher Inhalte sollten die Transparenzberichte von YouTube aussagekräftige Informationen über das Zusammenspiel von Maßnahmen zur Inhaltsmoderation und Empfehlungssystemen liefern. Dazu sollten granulare Daten gehören, die Forschern dabei helfen können, zu verstehen, wie Empfehlungssysteme zur Verbreitung schädlicher Inhalte beitragen können, und unabhängig zu überprüfen, ob die Schritte, die YouTube unternimmt, um Empfehlungen schädlicher Inhalte zu reduzieren, tatsächlich so funktionieren, wie das Unternehmen angibt.

YouTube bietet heute keine Transparenz darüber, wie es grenzwertige Inhalte definiert und behandelt. YouTube muss sich dieser Lücke in der Transparenz annehmen und sie schließen. YouTube sollte seine Transparenzberichterstattung um Informationen darüber erweitern, wie die Plattform „grenzwertige Inhalte“ definiert, welche Methoden der Inhaltsmoderation auf solche Inhalte anwendet werden (z. B. Herabstufung, Depriorisierung), und aggregierte Daten bereitstellen, die dabei helfen können, die Probleme im Zusammenhang mit dieser Kategorie von Inhalten auf der Plattform zu bewerten (z. B. wie oft YouTube grenzwertige Inhalte empfiehlt und wie groß die Gesamtmenge solcher Inhalte auf der Plattform ist).

Wichtig ist, dass diese Informationen auf einer ausgewiesenen Seite zugänglich sind – und nicht einfach in den Nutzungsbedingungen versteckt werden –, dass sie umfassend sind und in einer verständlichen Weise kommuniziert werden. Und um regionale Unterschiede zu überwachen, sollten diese Informationen nach Land/Geografie und Sprache aufgeschlüsselt werden.

3. **Nutzer:innen brauchen mehr Kontrolle darüber, wie ihre Daten als Input für die Generierung von Empfehlungen genutzt werden sowie den Output dieser Empfehlungen.**

   Plattformen sollten den Menschen mehr Kontrolle darüber geben, welche ihrer Daten zur Generierung von Empfehlungen verwendet werden. Die Nutzer:innen sollten auch vollen Einblick in und Kontrolle über andere Informationen haben, die bei der Erstellung von Empfehlungen berücksichtigt werden. Beispielsweise sollten Nutzer:innen die Möglichkeit haben, Daten auszuschließen, die von anderen verwandten Produkten/Diensten gesammelt wurden (z. B. Google-Daten, die für YouTube-Empfehlungen verwendet werden), von früheren Interaktionen mit bestimmten Inhalten/Seiten/Nutzer:innen, oder Daten über andere Personen, die zur Erstellung von Empfehlungen verwendet werden.

   Plattformen sollten es Menschen ermöglichen, die ihnen angezeigten Empfehlungen oder Inhalte anzupassen, um ihre Sicherheit auf der Plattform besser zu schützen, indem sie zusätzliche Kontrollen oder Auswahlmöglichkeiten bieten. Dies sollte z. B. die Möglichkeit beinhalten, bestimmte Schlüsselwörter, Arten von Inhalten oder Kanäle von Empfehlungen auszuschließen. Es sollte auch die Möglichkeit umfassen, zu kontrollieren, ob und in welchem Umfang

"grenzwertige Inhalte" oder bestimmte Kategorien von Inhalten auf der Plattform (z. B. Nachrichten, Sport) in Empfehlungen erscheinen.

Diese Steuerelemente sollten sowohl über zentrale Benutzereinstellungen zugänglich sein als auch in die Oberfläche integriert werden, die algorithmische Empfehlungen anzeigt. Wo es möglich ist, sollten Plattformen eine einfache Sprache verwenden, die das Ergebnis beschreibt, das sich aus der Verwendung eines Steuerelements ergibt, und nicht das Signal, das das Steuerelement sendet (z. B. „Zukünftige Empfehlungen von diesem Kanal blockieren“ anstelle von „Ich mag diese Empfehlung nicht“).

4. **Es sollten strenge, wiederkehrende Programme für das Risikomanagement eingerichtet werden, die sich ausschließlich mit Empfehlungssystemen befassen.**

   Plattformen sollten die Risiken für Einzelpersonen und das öffentliche Interesse, die sich aus der Gestaltung, der Funktionsweise oder der Nutzung des Empfehlungssystems ergeben können, systematisch und kontinuierlich ermitteln, bewerten und steuern. Eine solche Risikobewertung sollte sowohl die Wahrscheinlichkeit des Auftretens eines Schadens als auch das potenzielle Ausmaß des Schadens berücksichtigen. Durch einen derart umfassenderen Ansatz für das Risikomanagement wären Plattformen in der Lage, nicht nur die Geschäfts- und Reputationsrisiken, sondern auch die beabsichtigten und unbeabsichtigten externen Effekte, die ihre Dienste verursachen, besser zu erfassen.

5. **Nutzer:innen sollten die Personalisierung der ihnen angebotenen Inhalte deaktivieren können.**

   Plattformen, einschließlich YouTube, sollten Nutzer:innen die Möglichkeit bieten, personalisierte Empfehlungen abzulehnen und stattdessen chronologische, kontextbezogene oder rein suchbegriffsbasierte Empfehlungen zu erhalten. Der Zugriff auf den Dienst sollte nicht von der Anzeige von Empfehlungen (oder anderen Personalisierungsentscheidungen der Nutzer:innen) abhängig gemacht werden.

Unsere Empfehlungen an Gesetzgeber

1. **Fordern Sie YouTube und andere Plattformen dazu auf, Informationen freizugeben und Tools zu entwickeln, mit denen Forscher:innen die Empfehlungsalgorithmen der Plattformen mittels Audits und Datenzugriff unter die Lupe nehmen können.**

   Gesetzgeber sollten Regelungen einführen, die Plattformen zur Transparenz ihrer Empfehlungsalgorithmen verpflichten, sowie robuste Rahmenbedingungen für den Datenzugang, die eine unabhängige Untersuchung von Social-Media-Plattformen ermöglichen. Dies wird bereits im Vorschlag der Europäischen Kommission für ein Gesetz über digitale Dienste (Digital Services Act, DSA) vorgeschlagen und politische Entscheidungsträger in einer Reihe von anderen Ländern haben den Wunsch signalisiert, ähnliche Transparenz- und Aufsichtsregelungen einzuführen. Die politischen Entscheidungsträger müssen erkennen, dass YouTube und andere Plattformen diese dringend benötigte Transparenz nicht freiwillig bereitstellen und dass regulatorische Eingriffe notwendig sind.

2. **Es muss sichergestellt werden, dass die Rahmenbedingungen für die Gesetzgebung auf die einzigartigen Probleme und Risiken von Empfehlungsalgorithmen abstimmt werden.**

   Falls und wenn Rahmenbedingungen für die Verantwortung für Online-Inhalte geschaffen werden, sollten die Gesetzgeber sicherstellen, dass diese Regelungen Anreize für Vertrauen und Sicherheitsverantwortung für Systeme zur Empfehlung von Inhalten schaffen. Es gibt zahlreiche verschiedene Ansätze, wie dies erreicht werden könnte (z. B. durch die Verpflichtung zur Durchführung von Risikobewertungen oder das Angebot größerer Endnutzerkontrollen), aber grundsätzlich sollten Regulierungen, die die Verantwortung für Online-Inhalte betreffen, sicherstellen, dass Plattformen die Risiken bei der Entwicklung und dem Betrieb automatisierter Systeme, die Inhalte in großem Umfang verbreiten, angemessen berücksichtigen.

3. **Forscher:innen, Journalist:innen und andere Überwacher:innen, die alternative Methoden nutzen als die durch die Plattformen bereitgestellten, um diese zu untersuchen, müssen geschützt werden.**

Die politischen Entscheidungsträger sollten Safe-Harbor-Bestimmungen oder andere Schutzmaßnahmen schaffen, die Forscher:innen, die unabhängig von den von Plattformen bereitgestellten Kanälen für den Datenzugriff Forschung im öffentlichen Interesse betreiben, vor rechtlichen Bedrohungen schützen. Dies könnte verhindern, dass Plattformen bei der Gewährung des Zugangs zu Daten zu sparsam vorgehen. Zu diesem Zweck sollte die bestehende Gesetzgebung geändert oder klarer gefasst werden. So besteht beispielsweise nach wie vor erhebliche Unsicherheit hinsichtlich des Umfangs der in der EU-Datenschutzgrundverordnung (GDPR) vorgesehenen Ausnahmen für die Forschung; und trotz des kürzlich gefallenen ermutigenden Van-Buren-Urteils des Obersten Gerichtshofs der USA bleiben einige Fragen offen, wie der U.S. Computer Fraud and Abuse Act (CFAA) auf die Online-Forschung im öffentlichen Interesse anzuwenden ist. Um diese Unsicherheiten zu beseitigen, könnten Safe-Harbor-Bestimmungen Plattformen daran hindern, für die Forschung genutzte Tools zu blockieren oder Tarifbeschränkungen aufzuerlegen. Bestimmungen könnten einen besseren Schutz für Forscher:innen bieten, die Daten von Plattformen auslesen oder Sockenpuppen-Audits durchführen, wenn dies derzeit gegen die Nutzungsbedingungen der Plattformen verstößt.

Unsere Empfehlungen an Nutzer:innen von YouTube

1. **Informieren Sie sich darüber, wie YouTube-Empfehlungen funktionieren.**

   Sehen Sie sich Mozillas Videoserie „Mozilla Explains: Recommendation Engines“ und „Mozilla Explains: Recommendation Engines part 2“ an und lernen Sie im Detail, wie YouTube und andere Plattformen und deren Algorithmen für Empfehlungen funktionieren.

2. **Prüfen Sie Ihre Dateneinstellungen auf YouTube und Google und stellen Sie sicher, dass Sie die richtigen Kontrollen für sich und Ihre Familie aktiviert haben.**

   Viele Menschen sind sich der Existenz dieser Einstellungen nicht bewusst, da sie von YouTubes Homepage aus schwer zu finden sein können. Unser Profitipp ist, sicherzustellen, dass Sie Ihre Wiedergabe und Suchverläufe überprüfen und die

Videos daraus löschen, die Ihre Empfehlungen nicht beeinflussen sollen. Alternativ können Sie dies ganz ausschalten, was besonders hilfreich sein kann, wenn Sie sich einen Computer mit anderen Nutzer:innen teilen. Sie können auch diese Seite besuchen und lernen, wie Sie die automatische Wiedergabe auf YouTube von Ihrem Smartphone, TV oder Computer aus deaktivieren können.

3. **Laden Sie RegretsReporter herunter und stellen Sie Ihre Daten unserer durch Crowdsourcing angetriebenen Forschung zur Verfügung!**

   Mozilla wird RegretsReporter weiterhin als unabhängiges Tool zur Untersuchung des Empfehlungsalgorithmus von YouTube anbieten. Wir planen, die Erweiterung zu aktualisieren, um Menschen einen einfacheren Zugang zu den Nutzerkontrollen von YouTube zu ermöglichen und unerwünschte Empfehlungen zu blockieren, und wir werden weiterhin Analysen und Empfehlungen veröffentlichen.

Schlussfolgerung

YouTubes Algorithmus nimmt großen Einfluss darauf, was wir glauben – sowohl als Individuen als auch als Gesellschaft. Untersuchungen von Mozilla und zahllosen anderen Expert:innen haben bestätigt, dass erhebliche Risiken mit YouTubes Algorithmen einhergehen. Sollten diese Erfahrungen so häufig vorkommen, wie unsere Untersuchungen es vermuten lassen und sich auf andere KI-Systeme übertragen lassen (was in Anbetracht der kommerziellen Angebote wie Googles Empfehlungs-KI wahrscheinlich ist), werden die Auswirkungen enorm sein.

Viele Expert:innen sind der Meinung, dass die Probleme nicht auf Fehler im Algorithmus zurückzuführen sind, sondern vielmehr das Ergebnis des YouTube-Algorithmus sind, der genau so funktioniert, wie er soll. Das Problem liegt ihrer Meinung nach darin, dass es eine fundamentale Diskrepanz zwischen Algorithmen gibt, die für die Förderung von Geschäftsanreizen optimiert sind, und solchen, die für das Wohlergehen der Menschen optimiert sind. Das mag durchaus wahr sein. Was definitiv wahr ist, ist, dass Algorithmen, die so folgenreich sind, nicht ohne angemessene Aufsicht eingesetzt werden sollten. Und Transparenz ist ein wichtiger erster Schritt.

„Grenzwertige“ Inhalte sind subjektiv und schwer zu regulieren, egal ob sie codiert sind oder in der öffentlichen Gesetzgebung festgelegt – das liegt in ihrer Natur. Aber gerade weil es so schwierig ist, müssen sie untersucht und diskutiert und der Versuch unternommen werden, sie zu verstehen. Wir konnten unsere Untersuchung nur durchführen, weil wir die Fähigkeiten und Ressourcen hatten, ein Tool zu erstellen, das uns dies ermöglicht – das sollte nicht die Bedingung sein. YouTube könnte sich zur Offenheit bekennen und Forschung wie unsere ermöglichen, aber das Unternehmen hat sich dagegen entschieden. Stattdessen hat die Plattform sich dazu entschieden, ein Kartenhaus zu bauen und diese einflussreichen Entscheidungen ganz allein zu treffen. Communitys weltweit sind von den Entscheidungen betroffen, die YouTube in den Vorstandsräumen im Silicon Valley trifft, und verdienen Einsicht und Einfluss darauf.

YouTube hat deutlich gemacht, dass es den falschen Ansatz wählt, um mit dieser Verantwortung umzugehen. Wir werden dem richtigen Ansatz nur mit mehr Offenheit, Transparenz, Verantwortlichkeit und Demut näher kommen.

Methodik

Forschungsfragen

Unsere Forschung begann mit folgenden Fragen:

1. **Welche Videos werden als Regret gemeldet?**
2. **In welche Kategorien sind gemeldete Regrets einzuordnen?**
3. Gibt es erkennbare Muster in Bezug auf die Häufigkeit oder den Schweregrad gemeldeter Regrets?
4. Führen bestimmte Nutzungsmuster bei YouTube dazu, das vermehrt Inhalte empfohlen werden, die als Regret gemeldet werden?
5. **Gibt es geografische Unterschiede bei Regret-Quoten oder -Arten?**
6. **Wie häufig kommen Meldungen von Regrets vor und wie unterscheidet sich die Häufigkeit abhängig davon, wie zu den Inhalten gefunden wurde?**

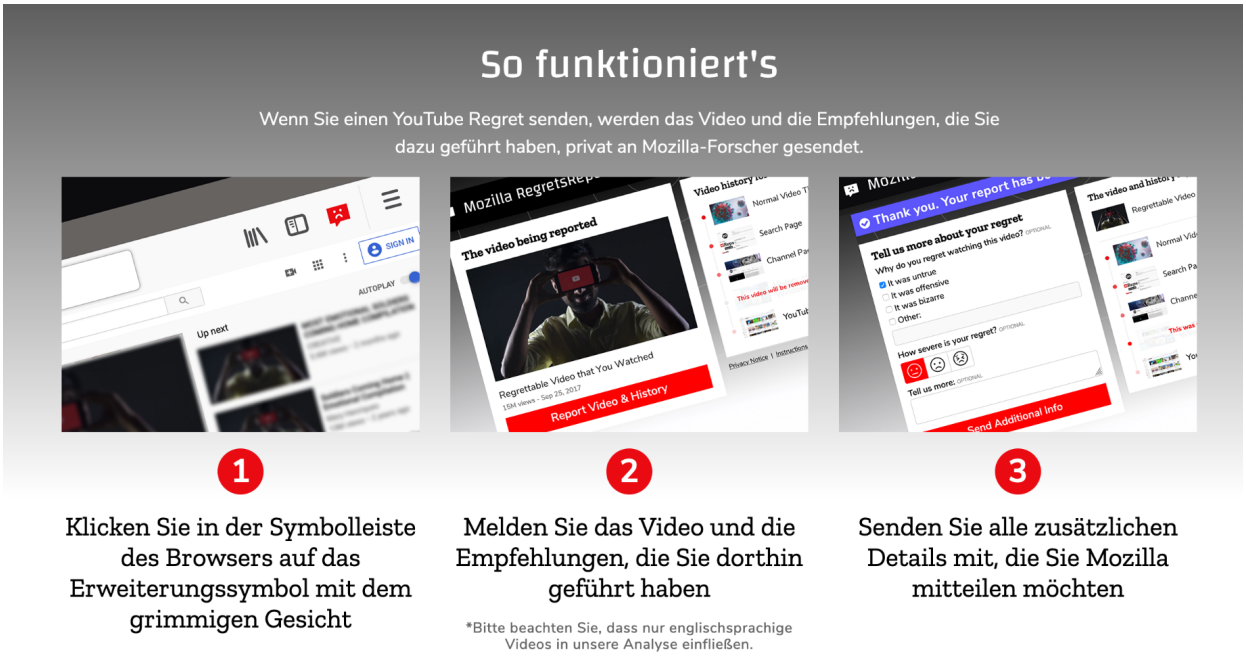
7. Ändert sich die Häufigkeit von Meldungen bei Nutzer:innen, wenn sie ihre erste Meldung getätigt haben? Wie unterscheidet sie sich nach Kategorie und Schweregrad?

8. **Was sind die qualitativen Merkmale der YouTube Regrets und des jüngeren Wiedergabeverlaufs, der den Meldungen vorausgeht? Gibt es erkennbare Muster?**

Schlussendlich konnten wir anhand unserer Datengruppe Fragen 1, 2, 5, 6 und 8 beantworten (fett gedruckt).

RegretsReporter-Erweiterung

RegretsReporter ist eine Browser-Erweiterung, die für Firefox und Chrome verfügbar ist und Freiwillige dazu befähigt, Daten über YouTube Regrets an Mozilla zu senden. Wer RegretsReporter installiert hat, kann das RegretsReporter-Icon in der Browserleiste einer YouTube-Videoseite klicken, um ein Meldeformular zu generieren. Das Meldeformular stellt eine Reihe an Fragen über die Erfahrung der Nutzer:innen und fasst die Daten zusammen, die an Mozilla gesendet werden.



Wird ein solcher Bericht gesendet, erhält Mozilla Informationen über das gemeldete Video inklusive Titel, Beschreibung, Anzahl der Aufrufe, Zugangspunkt (Empfehlungen, Suche etc.) und einen Link zur Seite des Videos sowie die optional von dem/der Freiwilligen bereitgestellten Daten wie Kategorie, Kommentar und Schweregrad der gemeldeten Inhalte. So können wir die gemeldete Erfahrung nachvollziehen und uns ein Bild davon machen, welche Art von Inhalten als Regret eingestuft werden.

Mozilla kann auch einen „Pfad“ dazu erhalten, wie der/die Freiwillige zu dem gemeldeten Video gekommen ist, wenn der/die Freiwillige sich entscheidet, diese Informationen zu senden. Der Pfad umfasst maximal die letzten fünf Seiten, die auf YouTube geladen wurden, und zwar höchstens innerhalb der letzten fünf Stunden vor der Meldung. Diese Informationen ermöglichen es uns, die Kette von Empfehlungen oder anderen Aktionen zu analysieren, die den:die Freiwillige:n zu dem gemeldeten Video geführt haben. Um die Quoten oder die Häufigkeit von Regrets zu berechnen, erhält RegretsReporter auch Daten über die YouTube-Nutzung jedes:jeder Freiwilligen, wobei keine Details darüber aufgezeichnet werden, welche Videos angeschaut werden oder wann genau YouTube genutzt wird, sondern ein Gesamtmaß dafür aufgezeichnet wird, wie viele Videos angeschaut werden. Diese Nutzungsdaten sind wichtig, um die Regret-Quoten aufgeschlüsselt nach verschiedenen Faktoren zu berechnen. Zu Guter Letzt enthalten alle Daten, die an Mozilla gesendet werden, das Land, aus dem jede:r Freiwillige YouTube besucht, basierend auf der GeoIP-Suche. Das erlaubt uns, zu analysieren, wie die Regret-Quoten zwischen den Ländern variieren.

RegretsReporter ist eine Open-Source-Software. Das bedeutet, dass jede:r auf den Code zugreifen kann. RegretsReporter wurde nach den Lean-Data-Prinzipien von Mozilla entwickelt, was bedeutet, dass **nur** Daten gesammelt werden, die zum Erreichen unserer Forschungsziele notwendig sind. RegretsReporter sammelt und speichert keine persönlichen Daten über die Freiwilligen, die es benutzen.

Ein von Menschen angetriebener Datensatz

Da YouTube den Forscher:innen keine Daten zur Verfügung stellt, entwickelten wir einen Citizen-Science-Ansatz, um die für die Beantwortung unserer Forschungsfragen benötigten Daten zu sammeln. Aufbauend auf und inspiriert von früheren Untersuchungen ist RegretsReporter die bisher größte Crowdsourcing-Untersuchung des YouTube-Empfehlungssystems. Unser Datensatz wird von 37.380 Freiwilligen aus 190 Ländern gespeist, die die RegretsReporter-Browser-Erweiterungen für Firefox und

Chrome installiert haben. Von unseren gesamten freiwilligen Teilnehmer:innen haben 1.662 mindestens einen Bericht eingereicht. Insgesamt waren es 3.362 Berichte aus 91 Ländern, die zwischen Juli 2020 und Mai 2021 eingereicht wurden. Freiwillige, die die Erweiterung heruntergeladen, aber keinen Bericht eingereicht haben, waren dennoch ein wichtiger Teil unserer Studie. Ihre Daten – z. B. wie oft sie YouTube nutzen – waren wesentlich für unser Verständnis, wie häufig YouTube[3] Regrets vorkommen und wie dies zwischen den Ländern variiert.

Dieser von Menschen angetriebene Ansatz fängt die realen Erfahrungen von Menschen ein, die YouTube nutzen, und ermöglicht uns trotz der mangelnden Bereitschaft von YouTube, Forscher:innen Daten zur Verfügung zu stellen, einen gewissen Einblick in den Algorithmus. Allerdings gibt es methodische Einschränkungen bei unserem Ansatz, darunter:

- Auswahlverzerrung: Unsere Proband:innen sind eine bestimmte Gruppe von Personen und unsere Ergebnisse lassen sich möglicherweise nicht auf alle YouTube-Nutzer:innen verallgemeinern.
- Berichtsverzerrung: Es kann viele Faktoren geben, die beeinflussen, ob ein:e Freiwillige:r ein bestimmtes Video meldet.
- Regret-Konzept: Das Konzept eines YouTube Regret ist (absichtlich) unspezifisch und Freiwillige haben möglicherweise unterschiedliche Grundsätze dafür, was sie als Regret einstufen und deshalb melden.
- Der beobachtende Charakter der Studie bedeutet, dass wir zwar mit Sicherheit sagen können, „was“ passiert, aber nicht mit Sicherheit auf das „Warum“ schließen können. Wir wissen zum Beispiel nicht, warum YouTube einem:einer bestimmtem:bestimmten Probanden:Probandin ein bestimmtes Video empfohlen hat.

Trotz dieser Einschränkungen bieten unsere Ergebnisse einen Einblick in Probleme auf YouTube aus der gelebten Perspektive echter Menschen aus der ganzen Welt. Wir glauben, dass die Einschränkungen unserer Forschung die Notwendigkeit dafür unterstreichen, dass YouTube den Forscher:innen Daten zur Verfügung stellt.

[3] Die Berechnung der Häufigkeit von Regrets hängt sowohl von der Anzahl der erlebten Regrets als auch von der Gesamtzahl der angesehenen Videos ab – 1 Regret unter 5 angesehenen Videos unterscheidet sich stark von 1 Regret unter 100 angesehenen Videos. Daher ist die Anzahl der von unseren Freiwilligen angesehenen Videos entscheidend, auch bei Freiwilligen, die keine Berichte erstellt haben.

Analysemethoden

Die Erweiterung RegretsReporter überträgt die gesammelten Daten über das Telemetriesystem von Mozilla und speichert sie anschließend im Data Warehouse von Mozilla. Die Analyse wurde mit BigQuery und Python hauptsächlich in der Google Colab-Umgebung durchgeführt. Der verwendete Analysecode ist hier öffentlich einsehbar.

Wir haben bei unserer Analyse mehrere verschiedene Methoden angewandt und dabei Schreibtischstudien, statistische Methoden und qualitative Analysen eingesetzt, um unsere Forschungsfragen zu beantworten.

Statistische Methoden umfassten die Berechnung von Anzahlen, Proportionen und Quoten, wobei nach relevanten Variablen unterschieden wurde. Inferentielle Techniken wurden angewendet, einschließlich der Berechnung von Konfidenzintervallen und der Auswertung von Hypothesentests. Alle in diesem Bericht beschriebenen Unterschiede waren auf dem Niveau $p < 0,05$ statistisch signifikant und alle statistischen Signifikanzen wurden auf diesem Niveau bewertet.

Wir begannen den Prozess der qualitativen Analyse mit der Einberufung einer Arbeitsgruppe von Expert:innen (die im Abschnitt „Danksagung“ dieses Berichts einzeln genannt werden), die über Erfahrungen in den Bereichen Online-Risiken, Meinungsfreiheit und Tech-Richtlinien verfügen. Die Arbeitsgruppe wurde damit beauftragt, die gemeldeten Videos aus dem Datensatz zu sichten und anschließend Themen zu identifizieren. Im Laufe von drei Monaten entwickelte die Arbeitsgruppe einen konzeptionellen Rahmen für die Klassifizierung einiger der Videos, basierend auf den YouTube-Community-Richtlinien. Die Arbeitsgruppe entschied sich, die YouTube-Community-Richtlinien als Leitfaden für die qualitative Analyse zu verwenden, da sie eine nützliche Systematik problematischer Videoinhalte bieten und auch eine Verpflichtung von YouTube darstellen, welche Art von Inhalten auf ihrer Plattform aufgenommen werden sollte.

Ein Team aus 41 wissenschaftlichen Mitarbeiter:innen (die im Abschnitt „Danksagung“ dieses Berichts einzeln genannt werden) von der University of Exeter nutzten diesen konzeptionellen Rahmen, um jedes gemeldete Video zu analysieren. Die wissenschaftlichen Mitarbeiter:innen sollten Fragen beantworten, wie beispielsweise:

- Sollte dieses Video Ihrer Meinung nach auf YouTube sein?

- Sollte dieses Video Ihrer Meinung nach unter „als nächstes abspielen" oder auf der YouTube Homepage empfohlen werden?
- Wenn Sie eine der obigen Fragen mit NEIN beantwortet haben: Gegen welche Kategorie der Community-Richtlinien könnte das Video Ihrer Meinung nach verstoßen?

Die wissenschaftlichen Mitarbeiter:innen analysierten auch die qualitativen Merkmale der Empfehlungspfade, die zu dem Regret-Video führten (sofern verfügbar), und beantworteten zusätzliche Fragen wie die Primärsprache des Videos und ob es um die COVID-19-Pandemie ging, was anderen Elementen unserer Analyse half.

Im Zeitraum vom 7.–15. Juni analysierten die wissenschaftlichen Mitarbeiter:innen 1141 oder 33,9 % aller 3361 Berichte.

Die Berichtsdaten wurden ab dem 22. Juli 2020 erhoben, zunächst mit einer ausgewählten Beta-Testgruppe, gefolgt von der Öffnung für die Teilnahme durch die Allgemeinheit im September 2020. Nur Daten, die bis zum 31. Mai 2021 eingingen, wurden für die Analyse berücksichtigt. Es wurden zwei Datenbereinigungsverfahren angewandt:

Inaktive Kriterien. Einige Freiwillige ließen die Erweiterung weiterhin installiert, nutzten aber die Berichtsfunktion nicht. Da die Absicht bei der Installation der Erweiterungen darin bestand, Berichte zu erstellen, betrachten wir diese Freiwilligen als inaktiv, wenn sie in einem Zeitraum von 56 Tagen (zwei 28-Tage-Monate) nach ihrem letzten Bericht bzw. seit der Installation keine Berichte erstellt haben. Nach Ablauf dieses 56-Tage-Zeitraums werden die Nutzungsdaten des:der Freiwilligen nicht mehr gezählt. Wenn der:die Freiwillige inaktiv wird und dann einen Bericht erstellt, wird er reaktiviert und seine:ihre vollen Nutzungsdaten werden gezählt.

Wir halten diesen Zeitraum für angemessen, da 91 % der Freiwilligen, die eine Meldung gemacht haben, dies innerhalb der ersten 56 Tage nach der Installation von RegretsReporter getan haben.

Ausreißerkriterien. Zwei Freiwillige scheinen die Erweiterung auf eine Art und Weise genutzt zu haben, die nicht mit dem Durchschnitt unserer Freiwilligen übereinstimmt. Diese beiden Freiwilligen haben 231 bzw. 109 Berichte eingereicht, während die nächstfolgenden Freiwilligen 65, 53 bzw. 37 Berichte eingereicht haben. Wir waren der

Meinung, dass die Einbeziehung der Daten dieser beiden Freiwilligen die Verallgemeinerbarkeit unserer Ergebnisse beeinträchtigen würde, und haben sie daher von der Analyse ausgeschlossen.

Offenlegungen

Wir sind fest entschlossen, klimaneutral zu sein und werden unseren Treibhausgas-Fußabdruck Jahr für Jahr deutlich reduzieren, um die Netto-Null-Emissionsverpflichtung des Pariser Klimaabkommens zu erfüllen und zu übertreffen. Wir nutzen Google Cloud Platform (GCP), um RegretsReporter zu hosten und Anfragen zu verarbeiten. GCP ist zu 100 % klimaneutral und RegretsReporter nutzt hauptsächlich die GCP-Infrastruktur in Oregon, die zu 89 % kohlenstofffreie Energie verwendet. Der volle Umfang der mit RegretsReporter verbundenen Kohlenstoffemissionen wird in Mozillas nächstem Greenhouse Gas (GHG) Inventory, das 2022 veröffentlicht werden soll, berechnet und offengelegt.

Alle für dieses Projekt gesammelten und verarbeiteten Daten wurden entweder von Mozilla-Mitarbeiter:innen oder von Forscher:innen bearbeitet, die vertraglich an die Datenschutz-, Sicherheits-, Datenverarbeitungs- und Vertraulichkeitsrichtlinien von Mozilla gebunden sind.

Wir erkennen den Umfang der Forschung zu den psychologischen Auswirkungen der Inhaltsmoderation und der damit verbundenen Arbeit an. Das Team, das an dieser Forschung gearbeitet hat, einschließlich der wissenschaftlichen Mitarbeiter:innen der University of Exeter, die sich den Großteil der gemeldeten Videos angesehen haben, um unsere Analyse zu unterstützen, wurde im Verlauf dieser Forschung psychologisch unterstützt.

Google ist in Firefox in vielen Regionen der Welt die Standardsuchmaschine. Trotz dieser Beziehung setzt sich Mozilla seit Langem für mehr Transparenz und Rechenschaft durch YouTube und andere Online-Plattformen ein.

Quellenangaben:

Adamczyk, Roman. “What’s the Hold-up? How YouTube’s inaction allowed the spread of a major French COVID-19 conspiracy documentary.” *EU DisinfoLab*, 12 May 2021, https://www.disinfo.eu/publications/whats-the-hold-up%3F-how-youtubes-inaction-allowed-the-spread-of-a-major-french-covid-19-conspiracy-documentary/

Alexander, Julia. “YouTube claims its crackdown on borderline content is actually working.” *The Verge*, 3 Dec. 2019, https://www.theverge.com/2019/12/3/20992018/youtube-borderline-content-recommendation-algorithm-news-authoritative-sources

Bergen, Mark. “YouTube Executives Ignored Warnings, Letting Toxic Videos Run Rampant.” *Bloomberg*, 2 Apr. 2019, https://www.bloomberg.com/news/features/2019-04-02/youtube-executives-ignored-warnings-letting-toxic-videos-run-rampant

Boyd, Ashley. “Senate Hearing Confirms YouTube Won’t Fully Release Recommendations Data Without More Pressure from Public and Congress.” *Mozilla*, 28 Apr. 2021, https://foundation.mozilla.org/en/blog/senate-hearing-confirms-youtube-wont-fully-release-recommendations-data-without-more-pressure-from-public-and-congress/

Bridle, James. “Something is wrong on the internet.” *Medium*, 6 Nov. 2017, https://medium.com/@jamesbridle/something-is-wrong-on-the-internet-c39c471271d2

Campbell, Eliza and Spandana Singh. “The flaws in the content moderation system: The Middle East case study.” *Middle East Institute*, 17 Nov. 2020, https://www.mei.edu/publications/flaws-content-moderation-system-middle-east-case-study

Chen, Annie Y., Brendan Nyhan, Jason Reifler, Ronald E. Robertson, and Christo Wilson. “Exposure to Alternative & Extremist Content on YouTube” *ADL*, Feb. 2021, https://www.adl.org/resources/reports/exposure-to-alternative-extremist-content-on-youtube

Cobbe, Jennifer and Jatinder Singh. “Regulating Recommending: Motivations, Considerations, and Principles.” *European Journal of Law and Technology (EJLT) 10(3)*, 30 Dec. 2019, https://ejlt.org/index.php/ejlt/article/view/686

Cook, Jesselyn and Sebastian Murdock. “YouTube Is A Pedophile’s Paradise.” *HuffPost*, 20 Mar. 2020, https://www.huffpost.com/entry/youtube-pedophile-paradise_n_5e5d79d1c5b6732f50e6b4db

Córdova, Yasodara, Adrian Rauchfleisch and Jonas Kaiser. “The implications of venturing down the rabbit hole.” *Internet Policy Review*, 27 Jun. 2019, https://policyreview.info/articles/news/implications-venturing-down-rabbit-hole/1406

Fisher, Max and Amanda Taub. “How YouTube Radicalized Brazil.” *The New York Times*, 11 Aug. 2019, https://www.nytimes.com/2019/08/11/world/americas/youtube-brazil.html

Fisher, Max and Amanda Taub. “On YouTube’s Digital Playground, an Open Gate for Pedophiles.” *The New York Times*, 3 Jun. 2019, https://www.nytimes.com/2019/06/03/world/americas/youtube-pedophiles.html

Geurkink, Brandi. “Congratulations, YouTube… Now Show Your Work.” *Mozilla*, 5 Dec. 2019, https://foundation.mozilla.org/en/blog/congratulations-youtube-now-show-your-work/

Geurkink, Brandi. “Our recommendation to YouTube.” *Mozilla*, 14 Oct. 2019, https://foundation.mozilla.org/en/blog/our-recommendation-youtube/

Leerssen, Paddy. “The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems.” *European Journal of Law and Technology (EJLT) 11(2)*, 31 Oct. 2020, http://www.ejlt.org/index.php/ejlt/article/view/786

Lewis, Rebecca. “Alternative Influence: Broadcasting the Reactionary Right on YouTube.” *Data & Society*, Sep. 2018, https://datasociety.net/wp-content/uploads/2018/09/DS_Alternative_Influence.pdf.

Maréchal, Nathalie and Ellery Roberts Biddle. “It's Not Just the Content, It's the Business Model: Democracy’s Online Speech Challenge.” *Ranking Digital Rights*, 17 Mar. 2021, https://rankingdigitalrights.org/its-the-business-model/

Molter, Vanessa. “Platforms of Babel: Inconsistent misinformation support in non-English languages.” *Election Integrity Partnership*, 21 Oct. 2020, https://www.eipartnership.net/policy-analysis/inconsistent-efforts-against-us-election-misinformation-in-non-english

Nicas, Jack. “How YouTube Drives People to the Internet’s Darkest Corners.” *The Wall Street Journal*, 7 Feb. 2018, https://www.wsj.com/articles/how-youtube-drives-viewers-to-the-internets-darkest-corners-1518020478

Ohlheiser, Abby. “They turn to Facebook and YouTube to find a cure for cancer — and get sucked into a world of bogus medicine.” *The Washington Post*, 25 Jun. 2019, https://www.washingtonpost.com/lifestyle/style/they-turn-to-facebook-and-youtube-to-find-a-cure-for-cancer--and-get-sucked-into-a-world-of-bogus-medicine

Ribeiro, Manoel Horta, Raphael Ottoni, Robert West, Virgílio A. F. Almeida, and Wagner Meira. “Auditing radicalization pathways on YouTube.” *In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)* (pp. 131–141), 27 Jan. 2020., https://doi.org/10.1145/3351095.3372879

Roberts, Sarah T. “Behind the Screen: Content Moderation in the Shadows of Social Media.” *Yale University Press*, 2019, https://doi.org/10.2307/j.ctvhrcz0v

Roth, Camille, Antoine Mazières and Telmo Menezes. “Tubes and bubbles topological confinement of YouTube recommendations.” *PLoS ONE 15(4)*, 21 Apr. 2020, https://doi.org/10.1371/journal.pone.0231703

Sanna, Leonardo, Salvatore Romano, Giulia Corona and Claudio Agosti. “YTTREX: Crowdsourced Analysis of YouTube’s Recommender System During COVID-19 Pandemic.” *Information Management and Big Data (pp.107-121)*, May 2021, https://doi.org/10.1007/978-3-030-76228-5_8

Singh, Spandana. “Why Am I Seeing This? How Video and E-Commerce Platforms Use Recommendation Systems to Shape User Experiences.” *Open Technology Institute*, 25 Mar. 2020, https://www.newamerica.org/oti/reports/why-am-i-seeing-this/

Singh, Spandana. “Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content.” *Open Technology Institute*, Jul. 2019, https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/

Tufekci, Zeynep. “YouTube, the Great Radicalizer.” *The New York Times*, 10 Mar. 2018, https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html

Vermeulen, Mathias. “The keys to the kingdom. Overcoming GDPR-concerns to unlock access to platform data for independent researchers”. *Knight First Amendment Institute Draft paper*, 27 Nov. 2020, https://doi.org/10.31219/osf.io/vnswz

“Answers to Your Questions About the Dark Side of the Internet.” *Mozilla*, 3 Sep. 2019, https://foundation.mozilla.org/en/blog/answers-your-questions-about-dark-side-internet/

“Continuing our work to improve recommendations on YouTube.” *YouTube Official Blog*, 25 Jan. 2019, https://blog.youtube/news-and-events/continuing-our-work-to-improve

“How YouTube Works.” *YouTube*, 2021. https://www.youtube.com/intl/howyoutubeworks/

“2020 Ranking Digital Rights Corporate Accountability Index.” *Ranking Digital Rights*, 2020, https://rankingdigitalrights.org/index2020/explore-indicators

“The Four Rs of Responsibility, Part 2: Raising authoritative content and reducing borderline content and harmful misinformation.” *YouTube Official Blog*, 3 Dec. 2019, https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-raise-and-reduce

“What Happened After My 13-Year-Old Son Joined the Alt-Right.” *Washingtonian*, 5 May 2019, https://www.washingtonian.com/2019/05/05/what-happened-after-my-13-year-old-son-joined-the-alt-right

Danksagungen

Dieser Bericht wurde von Jesse McCrosky und Brandi Geurkink geschrieben.

Mitwirkende Autoren waren: Kevin Zawacki, Anna Jay, Carys Afoko, Maximilian Gahntz und Owen Bennett.

Wir möchten uns bei den Mitgliedern unserer Arbeitsgruppe dafür bedanken, dass sie ihr Fachwissen mit uns geteilt haben, um den Rahmen für unsere Analyse zu verbessern: Gabrielle Guillemin, Dia Kayyali, Chico Camargo, Udbhav Tiwari, Jason Chuang und Amber Sinha. Die Ansichten in diesem Paper spiegeln nicht notwendigerweise ihre Ansichten oder die ihrer Arbeitgeber wider. Wir möchten auch den wissenschaftlichen Mitarbeiter:innen an der University of Exeter unter der Leitung von Dr. Chico Camargo für ihren Fleiß und ihre harte Arbeit bei der Analyse der gemeldeten Videos danken: Julia Dominika Burzyk, Maria Campbell, Giulia Catelani, Hoi Ying Chang, Sharon Choi, Hannah Cox, Isabel Dally, Ben Entwisle, Alice Gallagher Boyden, Laura Garratt, Adriano Giunta, Lisa Greghi, Rosemary Griggs, Matthew Gurney, Connie Hitchin, Olliver Hopkins, Oana Ionescu, Elliot Jones, Ritvika Kedia, Ruslan Kudryashov, Smriti Lakhotia, Michael Lewis, William Lewis, Mitran Malarvannan, Lois Mander, Zachary Marre, Alina McGregor, Inês Mendes de Souza, Ayodele Ogunyemi, Henry Payne, Sadaf Sahel, Matej Svoboda, Jia Tang Zhi, Martina Toneva, Olivia Warnes, Katherine Williams, Ching Yin Yan, Suchitra Bansode, Mingting Hong, Ayush Shrivastav, Feng Xu.

Wir danken unseren Kolleginnen und Kollegen aus der Zivilgesellschaft, die Beiträge, Hinweise und Rückmeldungen zu diesem Bericht geliefert haben, darunter Nathalie Maréchal, Spandana Singh und Mathias Vermeulen. Die Ansichten in diesem Paper spiegeln nicht unbedingt ihre Ansichten oder die ihrer Arbeitgeber wider.

Danke an Fred Wollsén, der RegretsReporter entwickelt hat, und an das Team von Reset Tech, die diese Arbeit finanziell unterstützt haben.

Abschließend möchten wir uns bei den 37.380 Freiwilligen bedanken, die über RegretsReporter Daten zur Verfügung gestellt haben. Unsere Lean-Data-Richtlinie bedeutet, dass wir nicht wissen, wer sie sind (daher können wir ihnen nicht einzeln danken), aber ohne ihre Beiträge wäre diese Forschung nicht möglich gewesen.

Anhang: Beispiele für YouTube Regrets nach Kategorie

Nachtragsbericht als PDF herunterladen